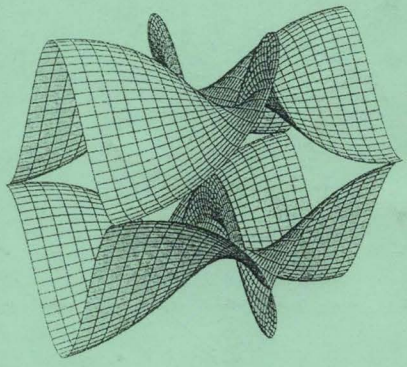
LUCIA MUTIHAC VASILE DAVID

ELEMENTE FUNDAMENTALE DE CHEMOMETRIE



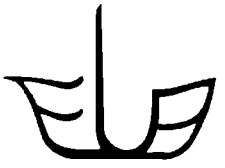
Editura Universităţii din Bucureşti

https://biblioteca-digitala.ro / https://unibuc.ro

https://biblioteca-digitala.ro / https://unibuc.ro

LUCIA MUTIHAC VASILE DAVID

ELEMENTE FUNDAMENTALE DE CHEMOMETRIE



EDITURA UNIVERSITĂȚII DIN BUCUREȘTI
2004

https://biblioteca-digitala.ro / https://unibuc.ro

4 Modelarea datelor analitice prin regresii mono sau multivariabile 49
4.1 Tipuri de regresii 49
4.2 Etapele analizei de regresie 53
4.2.1 Definirea modelului 53
4.2.2 Estimarea parametrilor modelului 57
4.2.3 Validarea modelului 63
4.2.4 Calculul intervalelor de încredere 67
4.3 Generalităţi privind determinările multicomponent 68
4.4 Concluzii 71
Bibliogarafie 71

5. Optimizarea şi proiectarea experimentelor 73
5.1 Aspecte ale optimizării în chimia analitică 73
5.1.1 Reprezentarea matematică a sistemului de optimizat 73
5.1.2 Funcţia de răspuns şi restricţiile 74
5.1.3 Strategia optimizării 77
5.1.4 Metode de evaluare a optimului funcţiei obiectiv 80
5.1.4.1 Metodele analitice 80
5.1.4.2 Metodele numerice 80
5.1.4.3 Metodele mixte 81
5.2 Metodologia suprafeţei răspunsului 83
5.2.1 Suprafaţa de răspuns 83
5.2.2 Experimente factoriale 87
5.2.2.1 Experiment factorial complet la două nivele 88
5.2.2.2 Experimente factoriale fracţionate 93
5.2.2.3. Experiment factorial la trei nivele 93
Bibliografie 94

6. Procesarea informaţiei chimice prin metodele inteligenţei artificiale 96
6.1 Reţele neuronale 96
6.1.2 Introducere 96
6.1.3 Noţiuni de teoria modelării 98
6.1.4 Modele de reţele neuronale 107
6.1.4.1 Reţele neuronale formale cu instruire supervizată (memorii asociative) 107
6.1.4.2 Reţele neuronale formale cu instruire nesupervizată (clasificatori vectoriali) 111
6.1.5 Algoritmi de instruire a reţelelor neuronale formale 113
6.1.5.1 Algoritmi de instruire generalizaţi 113
6.1.6 Aplicaţii ale reţelelor neuronale 120
6.1.6.1 Noţiuni fundamentale 120
6.1.6.2 Clase de aplicaţii specifice ale reţelelor neuronale 123

https://biblioteca-digitala.ro / https://unibuc.ro

6.1.6.3 Abordarea modelării neuronale 128
6.1.6.4 Aplicații ale rețelelor neuronale în chimie 130
6.1.6.5 Concluzii finale privind aplicațiile rețelelor neuronale în chimie 141
6.2 Algoritmi genetici 142
6.3 Analiza imagistică 144
Bibliografie 144

7 Procesarea semnalelor analitice 148
7.1 Tipuri de semnale 148
7.2 Eşantionarea semnalelor 151
7.3 Prelucrarea semnalelor analitice 153
7.3.1 Aspecte fundamentale ale prelucrării semnalelor 154
7.3.1.1 Caracterul temporal – frecvențial al semnalelor 154
7.3.1.2 Contaminarea cu zgomot a semnalelor reale 165
7.3.2 Tipuri de zgomot conținute în semnalele analitice 172
7.3.3 Tehnici de prelucrare a semnalelor analitice 174
7.3.3.1 Determinarea constantei de timp a instrumentelor analitice 174
7.3.3.2 Determinarea unor parametri fundamentali ai semnalelor 177
7.3.3.3 Tehnici de îmbunătățire a raportului semnal/zgomot 180
7.3.3.3.1 Tehnici instrumentale 180
7.3.3.3.1.1 Scurt istoric al utilizării tehnicilor de transformare în chimie 181
7.3.3.3.1.2 Sisteme analitice cu transformată Hadamard 182
7.3.3.3.2 Tehnici aplicate după achiziția semnalelor 185
7.4 Concluzii 187
Bibliografie 187

ANEXA 1 Analiza spectrală a semnalelor deterministe în domeniu de timp continuu 189
A1.1 Funcții ortogonale 189
A1.2 Serii Fourier 190
A1.3 Transformata Fourier continuă (CFT) 193

ANEXA 2. Planuri experimentale de cântărire 195
A2.1. Cazul clasic (al cântăririlor secvențiale) 195
A2.2. Cazul cântăririlor în grup punând obiectele de măsurat pe ambele talere 196
A2.3. Cazul cântăririlor în grup punând obiectele pe un singur taler 197
A2.4. Cazul general de cantarire în grup 198

https://biblioteca-digitala.ro / https://unibuc.ro

Lucrarea de față se adresează atât studenților cât și chimiștilor analiști preocupați de interpretarea rezultatelor experimentale cu ajutorul metodelor matematice moderne.

Lucia Mutihac

https://biblioteca-digitala.ro / https://unibuc.ro

1 Chemometria şi chimia analitică

1.1 Definirea chemometriei

Ca subdisciplină a chimiei, chemometria se ocupă cu aplicarea metodelor statistice şi matematice moderne în rezolvarea problemelor chimice în scopul obţinerii informaţiei optime. Aplicarea tehnicilor specifice tehnologiei informaţiei în optimizarea proceselor analitice şi evidenţierea direcţiilor fundamentale de evoluţie a metodelor computaţionale constituie alt aspect al obiectivelor chemometriei.

Dezvoltarea aparaturii instrumentale ca urmare a apariţiei tehnologiilor avansate a condus indirect la folosirea unor metode moderne de proiectare a experimentelor, calibrare a aparaturii instrumentale şi nu în ultimul rând la analiza datelor obţinute. Utilizarea metodelor computaţionale în studiul datelor chimice multivariate a fost posibilă în ultimii ani permiţând extragerea informaţiei pertinente. Astfel chemometria a crescut şi creşte în parallel cu dezvoltarea instrumentaţiei analitice precum şi cu posibilităţile computaţionale. Această combinaţie a deschis alte noi directii privind îmbunatăţirea prelucrării informaţiilor analitice.

Noţiunea de chemometrie a fost introdusă de suedezul Svante Wold şi americanul Bruce R. Kowalski în 1972. In chimia analitică, chemometria oferă metodele matematice şi statistice moderne în scopul obţinerii celor mai relevante informaţii asupra sistemelor materiale. Alături de metodele statistico-matematice, chemometria pune la îndemâna chimiştilor metodele inteligenţei artificiale în prelucrarea datelor chimice cum ar fi: reţelele neuronale, sisteme expert, analiza imagistică, algoritm genetic, etc. In general, reţelele neuronale sunt folosite în recunoaşterea de configuraţii. Sistemele expert sunt programme computerizate care ajută la rezolvarea problemelor la un nivel ştiinţific ridicat. Pe de altă parte, chemometria are un rol important în dezvoltarea altor metode cum ar fi metodele cromatografice de separare, automatizarea proceselor analitice sau în corelarea cantitativă structură-activitate a compuşilor chimici şi biologici.

https://biblioteca-digitala.ro / https://unibuc.ro

Chimia analitică, ca mod de investigare a unei ştiinţe experimentale, aparţine unei cercetări aplicate în care scopul este acela de a găsi o nouă metodă de analiză, de informare, de achiziţionare sau de îmbunătăţire a celor existente. În majoritatea cazurilor chimia analitică nu poate descoperi singură o lege ci ea trebuie să dispună de alte informaţii ca de exemplu informaţii fizice sau informaţii biologice.

Specificitatea chimiei analitice se poate înţelege dacă aceasta este abordată ca o stiinţă experimentală. Acest aspect este schematizat în jurul a 3 poli prezentaţi în figura 1.1 care constituie trinitatea analitică: obiectul studiat (o problemă), reprezentarea sa (un model) şi informaţia studiată.

Deci plecând de la o problemă, prima etapă constă în reprezentarea acesteia sub formă de model, în general matematic. Acest model serveşte la definirea naturii informaţiei cercetate şi organizează banca de date pe un plan teoretic. Originalitatea sa constă în informaţia cu caracterul chimic , ceea ce reprezintă o informaţie asupra materiei care cere o cunoaştere a concepţiilor de bază ale chimiei şi care se obţin cu ajutorul metodelor de măsurare chimică, fizică şi fizico-chimică puse în lucru de către analiza chimică. Rezultatele acestor analize vor fi structurate în acord cu modelul prestabilit şi evident se va lua decizia potrivită.

In concluzie, chimia analitică înglobează analiza chimică precum şi ansamblul metodelor de interpretare a rezultatelor în urma cărora se iau decizii cu scopul de a rezolva probleme propuse ca subiecte de studiu. Astfel, evaluarea sau controlul poluării unui râu în conţinutul de Pb este o problemă pusă chimiei analitice, şi în mod distinct, dozajul de Pb într-o probă dată este relevat de analiza chimică.

https://biblioteca-digitala.ro / https://unibuc.ro

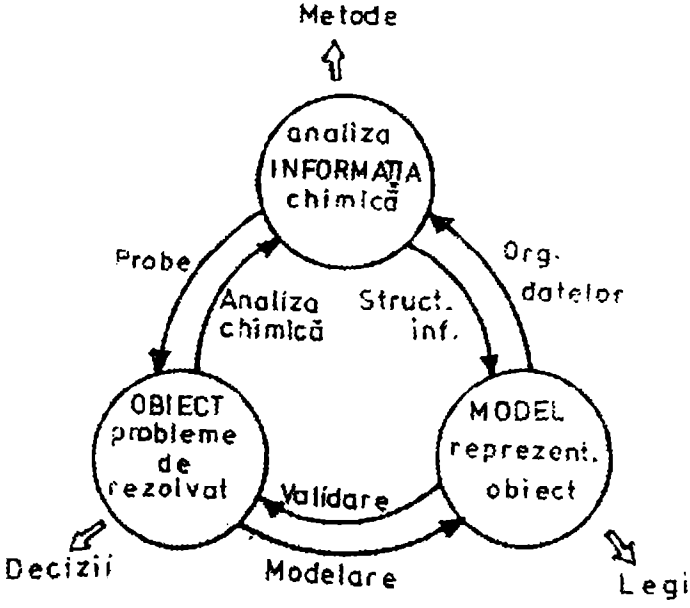


Figura 1.1 *Trinitatea analitică*

Chimia analitică este deci o metodologie care se bazează pe un ansamblu de metode chimice, fizice şi fizico-chimice, numite *metode de analiză chimică* şi metode informatice, statistice şi matematice adaptate cerinţelor specifice chimiştilor numite *metode chemometrice*. Ele au ca obiectiv rezolvarea problemelor prin intermediul procesării informaţiei chimice. De asemenea, rolul lor este de a optimiza şi îmbunătăţi colecţia de date utile, să se întrebe asupra conţinutului lor, asupra validităţii lor şi asupra semnificaţiei lor.

1.2 Obiectivele metodelor chemometrice

Începând cu anii 1970, atenţia mai multor grupuri de cercetători [1-6] a fost îndreptată spre elaborarea unor ansambluri de metode ale căror obiective încă de la apariţie se pot defini astfel:

- permit o mai bună prelucrare a rezultatelor obţinute de analişti în scopul extragerii informaţiei celei mai pertinente posibil;

https://biblioteca-digitala.ro / https://unibuc.ro

- creşterea calităţii măsurătorilor chimice prin optimizarea condiţiilor de colectare ale acestora.

S-au numit metode chemometrice. Chemometria este deci o disciplină încă tânără dar care a reuşit să se impună serios în chimie. In 1975 s-a creat *Chemometrics Society* iar în fiecare an are loc un congres intitulat *Chemometrics in Analytical Chemistry*. Există de asemenea două reviste de specialitate *Journal of Chemometrics* (Wiley) şi *Chemometrics and Intelligent Laboratory Systems* (Elsevier) unde sistematic sunt prezentate rezultatele obţinute în acest domeniu.

Se disting 4 domenii semnificative de investigaţie referitoare la posibilităţiile de aplicare a metodelor chemometrice şi anume:

a) achiziţionarea semnalelor la ieşirea din aparatele de măsură, stocarea lor şi aranjarea lor într-o formă corespunzătoare;

b) transformarea acestor semnale, expresia lor sub forma unor rezultate prin etalonare şi validarea lor în termeni de precizie;

c) interpretarea rezultatelor în vederea rezolvarii problemelor puse analiştilor;

d) gestionarea informaţiilor din laboratoarele de analiză.

Dezvoltarea chemometriei face parte integrantă din evoluţia chimiei. Chemometria a propus şi propune în continuare soluţii noi bazate pe un fundament matematic la problemele puse chimiştilor şi în principal, analiştilor [7-9]. Chemometria îşi are un loc bine definit în chimia analitică atât în specificitatea instrumentelor de măsură utilizate, natura probelor precum şi în modul de funcţionare al laboratorului. Calibrarea specrometrelor de infraroşu şi infraroşu apropiat este cea mai cunoscută aplicaţie a chemometriei atât în cercetare cât şi în industrie [10]. In ceea ce priveşte calculatorul într-adevăr acesta rămâne util cercetării şi dezvoltării, iar un obiectiv permanent îl constituie integrarea aparatelor la noile tehnici chemometrice.Tehnologia informatică a început să pătrundă în chimia analitică în jurul anilor 1960, în particular în tot ceea ce înseamnă instrumente de laborator. In consecinţă, majoritatea aparatelor de măsură disponibile astăzi conţin microprocesoare. Acestea vin să înlocuiască

https://biblioteca-digitala.ro / https://unibuc.ro

componentele electronice cablate ale aparatelor crescând astfel fiabilitatea instrumentelor şi îmbunătăţind în acelaşi timp funcţiile de filtrare ale semnalului, calculul statistic şi reprezentarea grafică. Dacă astăzi se vorbeşte de un instrument inteligent, această inteligenţă provine de la metodele informatice instalate în aparate (software specializat) şi nu de la microprocesorul însuşi.

Trăim într-o societate a informaţiei, chimia analitică este bogată în informaţii deci este logic ca dezvoltarea sa să beneficieze de metodele informaticii. Problema adaptării metodelor informaticii la cerinţele necesare analiştilor constituie unul dintre obiectivele principale ale chemometriei. Este important de a menţiona faptul că astăzi grupurile de cercetători chemometrişti colaborează cu bioinformaticieni, fizicieni, precum şi cu grupuri de de cercetători ai ingineriei calculatoarelor.

1.3 Informatica chimică

In ultimii 25 de ani au fost prezentaţi în literatura de specialitate sute de descriptori moleculari pornind de la indici topologici până la amprente ale complecşilor moleculari [11]. Este binecunoscut faptul că prelucrarea informaţiei chimice cu ajutorul calculatorului presupune reprezentarea moleculei prin descriptori moleculari care de cele mai multe ori necesită mii de *biţi* [12]. Noi tipuri de descriptori moleculari continuă să apară în acestă perioadă ca de exemplu descriptorul BCUT [13] pentru caracterizarea structurii moleculelor organice şi de asemenea descriptorul GRID [14,15] pentru caracterizarea structurii proteinelor. Descriptorii GRID pot fi uşor interpretaţi şi de asemenea computerizaţi. Reprezentarea structurii chimice este o problemă pentru informatica chimică. De asemenea o altă problemă o constituie explorarea unei cantităţi mari de date şi vizualizarea relaţiilor în baza de date chimice. Weinstein şi colaboratorii [16] au investigat diferite metode statistice şi metode ale inteligenţei artificiale pentru procesarea a 70 000 compuşi anticancer din baza de date cum ar fi: analiza componentului principal, analiza cluster, scalarea multidimensională şi reţelele neuronale. Cu ajutorul chimiei combinatoriale este posibil astăzi a se genera un număr mare de compuşi diverşi în baza de date. Aplicând metode statistice moderne unei baze de date conţinând mii de compuşi inclusiv caracteristicile lor biologice este

https://biblioteca-digitala.ro / https://unibuc.ro

posibilă creşterea vitezei de evaluare a acestora. Diversitatea bazei de date chimice este importantă în design-nul chimiei combinatoriale şi cea a experimentelor. Tehnica rezonanţei magnetice nucleare (RMN) s-a dovedit a fi o puternică şi eficientă metodă pentru monitorizarea interacţiilor agenţilor farmacologici cu celulele şi ţesuturile. Astfel folosind această tehnică (RMN) împreună cu recunoaşterea configuraţiilor (pattern recognition) se poate detecta toxicitatea indusă de unele medicamente [17].

Bibliografie

1. M. A. Sharaf, D. L. Illman, B. R. Kowalski, *Chemometrics*, John Wiley, New York, 1986.
2. D. L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, *Chemometrics : a textbook*, John Wiley, New York, 1988.
3. D. L. Massart, A. Dijkstra, L. Kaufman, *Evaluation and Optimization of Laboratory Methods and Analytical Procedures*, Elsevier, Amsterdam, 1978.
4. D. L. Massart, B.G.M. Vandeginste, L. M. C. Buydens, S. De Jong, P. J. Levi, J Smeyers-Verbeke, *Handbook of Chemometrics and Qualimetries, Part A*, Elsevier Science, Amsterdam, 1997.
5. G.Kateman, J. Pijpers, *Quality Control in Analytical Chemistry*, John Wiley, New York, 1981.
6. M. Feinberg, *L'Actualité Chimique* 3, 191, 1991.
7. M. Otto, *Chemometrics, Statistics and Computer Application in Analytical Chemistry*, J.Wiley-VCH, Weinheim, 1999.
8. L. Mutihac, V. David, *Chemometrie*, Ed. Universităţii din Bucureşti, 1997.
9. C. Bala, L. Mutihac, *Chemometrie, Principii şi aplicaţii în chimia analitică*, Ed. Universităţii din Bucureşti, 2000.
10. S. Wold, M. Josefsson, Multivariate Calibration of Analytical Chemistry, in *Encyclopedia of Analytical Chemistry*, A. R. Meyer, Ed., J. Wiley & Sons, New York, 9710-9736, 2000.
11. B. K. Lavine, J. Workman, Jr., *Anal. Chim. Acta* 74, 2763, 2002.
12. L. Xue, J. Bajorath, *Comb. Chem. High Throughput Screening* 3, 363, 2000.

https://biblioteca-digitala.ro / https://unibuc.ro

13. B. Pirard, S. D. Pickett, *J. Chem. Inf. Comput. Sci.* 40, 1431, 2000.

14. M. Pastor, G. Cruciani, I. McLay, S. Pickett, S. Clementi, *J. Med. Chem.* 43, 3233, 2000.

15. E. Fillipponi, V. Cecchetti, O. Tabarrini, D. Bonelli, A. Fravolini, *J. Comput.-Aided Mol. Des.* 14, 277, 2000.

16. L. Shi, Y. Fan, J. Lee, M. Waltham, D. Andrews, U. Scherf, K. Paulli, J. Weinstein, *J. Chem. Inf. Comput. Sci.* 40, 367, 2000.

17. E. Holmes, J. Shockcor, *Curr. Opin. Drug Discovery Dev.* 3, 72, 2000.

https://biblioteca-digitala.ro / https://unibuc.ro

2 Criterii utilizate în evaluarea metodelor analitice

2.1 Tipuri de erori în analiza chimică

Scopul determinărilor în chimia analitică îl constituie estimarea valorii adevărate. Dar datorită faptului că un proces analitic, în general, este supus unor surse de zgomot (perturbații) inerente, rezultatele măsurătorilor şi apoi evident rezultatele de analiză vor fi însoțite de erori. Cu cât concentrația componentului de determinat este mai mică, cu atât probabilitatea apariției erorilor în determinare este mai mare.

În concluzie, pentru ridicarea performanțelor unei determinări analitice este necesară îndepărtarea surselor de erori, lucru posibil evident prin cunoaşterea naturii acestora.

Eroarea de determinare analitică este dată de diferența dintre rezultatul determinării şi valoarea adevărată a determinării. Dacă se notează cu $\hat{\mathbf{x}}_O$ valoarea adevărată a conținutului componentului de determinat, iar cu $\mathbf{x}$ rezultatul analizei, eroarea de determinare se poate scrie:

$$\Delta\mathbf{x} = \mathbf{x} - \hat{\mathbf{x}}_O \qquad (2.1)$$

Eroarea relativă de determinare are următoarea expresie:

$$(\Delta\mathbf{x})_r = (\mathbf{x} - \hat{\mathbf{x}}_o)/\hat{\mathbf{x}}_o \qquad (2.2)$$

iar procentual în raport cu valoarea adevărată a conținutului, eroarea relativă de determinare se va exprima astfel:

$$(\Delta\mathbf{x})\% = \frac{(\mathbf{x} - \hat{\mathbf{x}}_o)}{\hat{\mathbf{x}}_o} \cdot 100 \qquad (2.3)$$

Clasificarea erorilor de determinare în mod frecvent cuprinde [1]: erori întâmplătoare şi erori sistematice. Între erorile întâmplătoare şi cele sistematice nu există o diferențiere rigidă şi anume în funcție de modul de examinare a erorilor, erorile sistematice pot trece în erori întâmplătoare şi invers. **Erorile întâmplătoare** se pot evidenția atunci când la repetarea unui experiment de determinare se obține de

https://biblioteca-digitala.ro / https://unibuc.ro

fiecare dată alt rezultat. În concluzie, eroarea întâmplătoare este dată de diferența dintre rezultatele obținute și valoarea medie a acestora la efectuarea repetată a unui experiment. Erorile întâmplătoare nu pot fi prevăzute individual, dar mulțimea lor poate fi descrisă prin funcții de repartiție bine determinate.

Erorile sistematice sunt acele erori de analiză care prezintă un caracter constant, adică distorsionează rezultatele determinărilor într-un singur sens cu o cantitate constantă față de valoarea adevărată. De asemenea este necesar, în funcție de stabilitatea sistemului analitic în timp, să fie luat în considerare și parametrul timp la clasificarea erorilor sistematice și anume există erori sistematice variabile în timp și erori sistematice constante în timp. Erorile sistematice apar de exemplu în cazul folosirii incorecte a instrumentelor de măsură datorită unei funcții de calibrare incorecte, folosirea unor reactivi de puritate necorespunzătoare, fenomene de interferență, etc.

Altfel exprimat, erorile sistematice se pot clasifica în dependente și independente de conținutul componentului de determinat.

2. 2 Precizia și acuratețea în chimia analitică

Precizia și acuratețea sunt două dintre cele mai importante criterii de caracterizare a sistemelor analitice precum și a rezultatelor acestora.

2.2.1 Precizia (reproductibilitatea) este dată de împrăștierea rezultatelor unei mulțimi de determinări în jurul mediei lor și este evaluată de abaterea standard s_x. Dacă $x_1, x_2, \ldots, x_n$ este mulțimea rezultatelor, n este numărul rezultatelor și $\bar{x}$ media rezultatelor, atunci dispersia de selecție s_x^2 are următoarea expresie:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad (2.4)$$

In cazul în care seria are valori care se repetă, expresia 2.4 va avea forma de mai jos:

https://biblioteca-digitala.ro / https://unibuc.ro

$$s_x^2 = \frac{\sum_{i=1}^{n} n_i (x_i - \bar{x})^2}{N} \qquad (2.5)$$

unde n_i este numărul valorilor x_i iar $N = \sum_{i=1}^{n} n_i$.

Abaterea standard sau abaterea medie pătratica de selecție s_x este rădăcina pătrată a dispersiei de selecție s_x^2 :

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \quad , \qquad (2.6)$$

respectiv:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} n_i (x_i - \bar{x})^2}{N}} . \qquad (2.7)$$

Rădăcina pătrată a dispersiei de selecție modificată este abaterea standard de selecție modificată:

$$s = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \qquad (2.8)$$

Aceasta este expresia abaterii standard folosită în practica statistică pentru estimarea abaterii standard a întregii populații [2].

În baza definiției abaterii standard rezultă că, *abaterea standard de estimație* pentru rezultatele medii a **n** determinări este:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}, \qquad (2.9)$$

abaterea standard relativă de estimație (coeficientul de variație) şi respectiv *abaterea standard relativ procentuală de estimație* au expresiile de mai jos:

$$RSD = \frac{s_x}{\bar{x}} \qquad \text{şi} \qquad \%RSD = \frac{s_x}{\bar{x}} \cdot 100 \qquad (2.10)$$

https://biblioteca-digitala.ro / https://unibuc.ro

deviația medie absolută de estimație este $\frac{\sum_{i=1}^{n} |\bar{x} - x_i|}{n}$.

Pentru a elimina orice eroare de interpretare, și de asemenea, pentru ca abaterea standard să aibă o semnificație tehnică, este necesară menționarea condițiilor în care a fost obținută mulțimea de date $x_1, x_2, \ldots, x_n$. În literatura de specialitate, pentru noțiunea de precizie se mai utilizează și următoarea nomenclatură: reproductibilitate, repetabilitate, împrăștiere, imprecizie [3-6].

Deci, evaluarea preciziei se face prin intermediul abaterii standard, s_x, cu specificarea obligatorie a condițiilor și modului de obținere a mulțimii de date $x_1, x_2, \ldots, x_n$.

Precizia este un criteriu de măsurare a reproductibilității măsurătorii. Dacă diferența între măsurătorile rezultate este mare, precizia este mică. O precizie bună nu înseamnă că și măsurătoarea este exactă deoarece este posibil să se comită sistematic aceeași eroare.

2.2 2 Acuratețea unei metode de analiză este dată de deviația rezultatelor simple sau a rezultatelor medii de la valoarea adevărată. Pentru evaluarea acurateții este necesar ca probele standard să aibă compoziție similară cu cea a probei de analizat, deoarece în majoritatea cazurilor acuratețea este influențată de compoziția probei. Acuratețea este afectată de eroarea sistematică [7-9].

2. 3 Selectivitatea și rezoluția

Problema definirii și evaluării selectivității în chimia analitică a preocupat si preocupă o mare parte din cercetătorii analiști astfel că această problemă a constituit un subiect de interes în literatura de specialitate [10-14]. Fiecare probă analitică este un amestec de mai mulți componenți și anume *analiții* care sunt de altfel componenții de interes care urmează a fi determinați și *matricea* reprezentată de ceilalți componenți prezenți în probă care nu prezintă interes din punctual de vedere al

https://biblioteca-digitala.ro / https://unibuc.ro

determinării respective, dar care pot influența măsurătorile. In chimia analitică acest efect este cunoscut sub denumirea de *efect de matrice*.

Selectivitatea este dată de posibilitatea unei metode de a detecta sau determina unul sau mai mulți componenți din probă în prezența altor componenți. Ea poate fi evaluată cu ajutorul unei funcții de răspuns, adică al funcției de dependență dintre răspuns (**y**) și compoziția probei de analizat. În cazul unei funcții de răspuns liniare în raport cu conținutul componenților se poate scrie următoarea relație:

$$\mathbf{y = a_1c_1 + a_2c_2 + ... + a_nc_n} \qquad (2.11)$$

unde $\mathbf{c_1}$ este conținutul componentului de analizat iar $\mathbf{c_2, c_3, ..., c_n}$ reprezintă conținutul celorlalți componenți. Metoda este selectivă dacă $\mathbf{a_1}$ este singurul coeficient diferit de zero.

Dacă se consideră cazul unui sistem cu **m** componenți și funcția de răspuns liniară în raport cu conținutul componenților, sistemul de **m** ecuații se poate scrie:

$$\begin{aligned} \mathbf{y_1} &= \mathbf{a_{11}c_1 + a_{21}c_2 + ... + a_{n1}c_n} \\ \mathbf{y_2} &= \mathbf{a_{12}c_1 + a_{22}c_2 + ... + a_{n2}c_n} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathbf{y_m} &= \mathbf{a_{1m}c_1 + a_{2m}c_2 + ... + a_{nm}c_n} \end{aligned} \qquad (2.12)$$

Matricea corespunzătoare acestui sistem de unde se poate extrage informația referitoare la selectivitatea sistemului are expresia de mai jos:

$$\begin{Vmatrix} \mathbf{a_{11}} & \mathbf{a_{21}} & \cdots & \mathbf{a_{n1}} \\ \mathbf{a_{12}} & \mathbf{a_{22}} & \cdots & \mathbf{a_{n2}} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{a_{1m}} & \mathbf{a_{2m}} & \cdots & \mathbf{a_{nm}} \end{Vmatrix} \qquad (2.13)$$

Se menționează faptul că sunt selective acele canale pentru care numai coeficientii de tip $\mathbf{a_{11}}$ sunt diferiți de zero. Dacă funcția de răspuns nu este liniară în

https://biblioteca-digitala.ro / https://unibuc.ro

raport cu conținutul componenților evident funcția de răspuns nu mai corespunde formei (2.11) iar matricea corespunzătoare nu are forma (2.13).

Tehnica cea mai potrivită pentru evaluarea selectivității unei metode este metoda experimentului factorial planificat însoțită de analiza de regresie.

Evaluarea selectivității unei metode nu este o problemă simplă. Ea presupune corelarea cunoştiințelor matematice şi statistice şi apoi adaptarea acestora la informația chimică cerută.

Influențarea răspunsului sistemului pentru conținutul unui component de către conținutul celorlalți componenți din probă cunoscut sub denumirea de *efect de matrice* după cum s-a menționat mai sus conduce la fenomenul de *interferență*.

Fenomenul de *interferență* se produce atunci când semnalul elementului de determinat nu poate fi separat de semnalele altor componenți din proba de analizat. Aceasta se întâmplă în cazul liniilor spectrale, undelor polarografice sau picurilor cromatografice. In spectrometria UV-Vis de exemplu există cazuri când mai mulți compuşi absorb la aceeaşi lungime de undă iar atunci când sunt prezenți în aceeaşi probă, absorbanța masurată este dată de suma absorbanților fiecărui component în parte. În acest caz, se foloseşte *rezoluția* ca măsură a gradului de suprapunere a două semnale experimentale. În figura 2.1 este prezentată grafic definirea rezoluției.

Gradul de suprapunere al celor două picuri este dat de raportul:

$$\mathbf{R_s = 2\frac{\Delta x}{x_1 + x_2}}. \qquad (2.14)$$

condiția pentru un grad de separare de 98% este prezentată în expresia 2.15:

$$\mathbf{\Delta x = 2\sigma_1 + 2\sigma_2} \qquad (2.15)$$

caz pentru care $\mathbf{R_s = 1}$

Măsura gradului de suprapunere a două semnale, de orice formă, este evaluată prin integrala:

$$\mathbf{A = \int_{-\infty}^{+\infty} y_1(x) y_2(x) dx} \qquad (2.16)$$

https://biblioteca-digitala.ro / https://unibuc.ro

în care $y_1(x)$ şi $y_2(x)$ reprezintă expresiile funcţiilor celor două semnale. Când cele doua semnale sunt separate, **A** = 0, iar când se suprapun **A>0**; cu cât gradul de suprapunere este mai mare, cu atât parametrul **A** are o valoare mai mare.

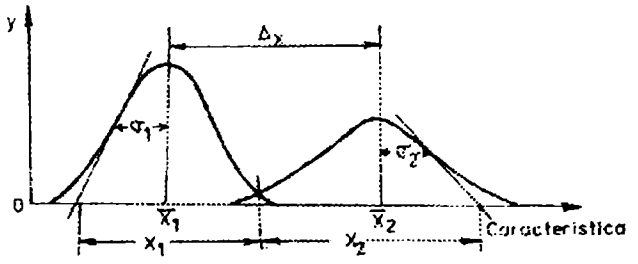


Figura 2.1 *Reprezentarea schematică a noţiunii de rezoluţie*

Fenomenul de *interinfluenţă* este dat de modificarea amplitudinii semnalului în funcţie de valorile conţinutului altor componenţi din probă. Acest fenomen se întâlneşte adesea în cazul spectroscopiei atomice de emisie şi absorbţie.

În chimia analitică, îmbunătăţirea selectivităţii sistemelor analitice se poate realiza fie prin perfecţionarea instrumentaţiei analitice, fie prin prelucrarea semnalului.

2.4 Sensibilitatea şi limita de detecţie

Adesea se spune despre o metodă că este foarte sensibilă atunci când limita de detecţie are o valoare foarte mică. Sensibilitatea, **S**, este definită ca fiind derivata funcţiei de răspuns $\mathbf{f}(\hat{\mathbf{x}})$ în raport cu conţinutul de determinat, şi anume:

$$S = \frac{\partial \mathbf{f}(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}}. \qquad (2.17)$$

În cazul unei funcţii de răspuns liniare,

$$y = a + bx \qquad (2.18)$$

sensibilitatea este egală cu panta funcţiei:

$$S = b = \frac{\Delta f}{\Delta x} \qquad (2.19)$$

https://biblioteca-digitala.ro / https://unibuc.ro

iar pentru o funcție de răspuns neliniară, sensibilitatea se modifică în raport cu valoarea conținutului de determinat. Pentru concentrații mici de analit, precizia determinărilor scade, iar în acest caz, sensibilitatea poate fi considerată o funcție de precizie. Se definește astfel, *sensibilitatea analitică* pentru o anumită concentrație de analit, ca fiind raportul dintre sensibilitate, *b* (panta dreptei în cazul unei curbe de calibrare liniare) și abaterea standard a semnalului, σ.

$$S = \frac{b}{\sigma} \qquad (2.20)$$

Sensibilitatea unei determinări analitice este constantă pentru un anumit domeniu de concentrații care trebuie specificat. Limita inferioară a domeniului de concentrații în care sensibilitatea poate fi considerată constantă, este dată de limita de detecție, iar limita superioara corespunde concentrației de analit pentru care sensibilitatea se modifică.

Caracterizarea capacității de detecție a unei metode de analiză constituie una dintre cele mai importante probleme ale chimiei analitice. Bazele acestei probleme au fost puse de Kaiser [15] și Wilson [16], Kaiser fiind acela care introduce noțiunea de *limită de detecție*. Limita de detecție, conform IUPAC-ului [17] este dată de concentrația sau cantitatea de analit care conduce la cel mai mic semnal analitic care poate fi măsurat cu o anumită *incertitudine* pentru o anumită procedură analitică. Morrison [18] definește limita de detecție ca fiind concentrația cea mai mică de analit care poate fi detectată cu certitudine într-un proces analitic. *Limita de determinare* poate fi definită ca limita pentru care o metodă analitică este suficient de precisă pentru a permite măsurarea valorii conținutului de determinat pentru analitul de interes sau concentrația care poate fi determinată cu o abatere relativă standard cu valoare cunoscută și maximă [19].

Contribuții însemnate în rezolvarea problemei sensibilității și a limitei de detecție aduc Ehrlich și Gerbatsch [20]. Liteanu și Rîcă [21] fundamentează definirea limitei de detecție pe baze statistice de detecție a semnalelor luând în considerare dependența dintre semnalul analitic și conținutul componentului de determinat.

https://biblioteca-digitala.ro / https://unibuc.ro

Semnalul analitic corespunzător limitei de detecție în sensul unui conținut minim sigur detectabil pentru un raport semnal/zgomot $\mathbf{r_d} = 6$, atunci când detecția se face cu un rezultat simplu, are următoarea expresie:

$$\hat{\mathbf{y}}_\mathbf{d} = \hat{\mathbf{y}}_\mathbf{0} + \mathbf{6}\sigma_\mathbf{y} \tag{2.21}$$

unde σ_y este *abaterea standard*, iar pentru media a **n** determinări simple avem expresia de mai jos:

$$(\hat{y}_d)_n = \hat{y}_0 + 6\frac{\sigma_y}{\sqrt{n}} \tag{2.22}$$

În cazul unei dependențe liniare între semnal şi conținut, este satisfăcută următoarea relație:

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_\mathbf{0} + \mathbf{b}\hat{\mathbf{c}}_\mathbf{d} \tag{2.23}$$

unde $\hat{\mathbf{c}}_\mathbf{d} = \mathbf{r_d}\dfrac{\sigma_\mathbf{y}}{\mathbf{b}}$, iar pentru condițiile de definiție alese $\hat{\mathbf{c}}_\mathbf{d} = \mathbf{6}\dfrac{\sigma_\mathbf{y}}{\mathbf{b}}$. Semnalul analitic corespunzător limitei de detecție în cazul mediei rezultatelor măsurătorilor are următoarea expresie:

$$(\hat{y}_d)_n = \hat{y}_0 + r_d\frac{\sigma_y}{\sqrt{n}} \tag{2.24}$$

În concluzie, între limita de detecție corespunzătoare rezultatelor medie a **n** rezultate simple şi limita de detecție pentru rezultate simple, este relația următoare:

$$(\hat{c}_d)_n = \frac{\hat{c}_d}{\sqrt{n}} \tag{2.25}$$

Liteanu şi Florea [22] au elaborat o metodă frecventometrică cu ajutorul careia se poate evalua riguros limita de detecție în sensul unui conținut minim detectabil. Long şi Winefordner [23] tratează metodele de determinare a limitei de detecție în funcție de erorile de măsură ale determinărilor analitice.

https://biblioteca-digitala.ro / https://unibuc.ro

2. 5 Amplitudinea semnalului şi raportul semnal/zgomot

În chimia analitică, rezultatul unei măsurători poate fi considerat suma următoarelor semnale: semnalul fondului, semnalul dat de componentul de determinat şi zgomotul ce însoţeşte orice proces de măsură. Dacă se reprezintă ipotetic o funcţie de răspuns pentru un sistem analitic, aceasta are forma $\hat{y} = \hat{y}_0 + f(c)$, în care $\hat{y}_0$ reprezintă semnalul fondului pentru o valoare $\hat{c}_0$ a conţinutului componentului de determinat care produce un semnal egal cu semnalul fondului. În acest caz, valoarea adevărată pentru semnalul $f(\hat{c}_i)$ dat de conţinutul $\hat{c}_i$ pentru componentul de determinat este dată de expresia:

$$f(\hat{c}_i) = \Delta\hat{y}_i = \hat{y}_i - \hat{y}_0 \qquad (2.26)$$

În cazul unei distribuţii normale a semnalului Δy_i, raportul semnal/zgomot corespunzător unei concentraţii c_i are următoarea expresie:

$$\frac{S}{N} = \frac{\hat{y}_i - \hat{y}_0}{\sigma_{\Delta y_i}} \qquad (2.27)$$

în care $\sigma_{\Delta y_i}$ este abaterea standard. Raportul semnal/zgomot este considerat caracteristică de performanţă în domeniul conţinuturilor mici; în domeniul conţinuturilor mari, semnificativ este inversul raportului semnal/zgomot, adică:

$$\frac{N}{S} = \frac{\sigma_{\Delta y_i}}{\hat{y}_i - \hat{y}_0} \qquad (2.28)$$

care reprezintă abaterea standard relativă pentru rezultatul măsurării semnalului conţinutului componentului de determinat.

Bibliografie

1. C. Liteanu, I. Rîcă, *Optimizarea proceselor analitice*, Ed. Academiei Române, Bucuresti, 1985.
2. D. Ceauşescu, *Utilizarea statisticii matematice în chimia analitică*, Ed. Tehnică, Bucureşti, 1982.

https://biblioteca-digitala.ro / https://unibuc.ro

3. C. Liteanu, I. Rîcă, *Teoria şi metodologia statistică a analizei urmelor*, Ed. Scrisul Românesc, Craiova, 1979.

4. D. L. Massart, A. Dijkstra, L. Kaufman, *Evaluation and Optimization of Laboratory Methods and Analytical Procedures*, Elsevier, Amsterdam,1978.

5. A. L. Wilson, *Talanta* 17, 31, 1970.

6. W. J. Younden, E. H. Steiner, *Statistical Manual of the Association of Official Analytical Chemists*, The Association of Official Analytical Chemists, Washington DC, 1975.

7. E. F. McFarren, R. J. Lishka, J. H. Parker, *Anal. Chem.* 42, 358, 1970.

8. K. Eckschlager, *Anal. Chem.* 44, 878, 1972.

9. K. Eckschlager, *Coll. Czech. Chem. Comm.* 39, 1426, 1974.

10. R. Belcher, *Talanta* 12, 129, 1965; *Talanta* 23, 883, 1976.

11. R. Belcher, D. Betteridge, *Talanta* 13, 535, 1966.

12. H. Kaiser, *Z. Analyt. Chem.* 260, 252, 1972.

13. A. L. Wilson, *Talanta* 12, 701, 1965; *Talanta* 21, 1109, 1974.

14. L. Pszonicki, *Talanta* 24, 613, 1977; *Chem. Anal.* 25, 439, 1980.

15. H. Kaiser, *Optik* 21, 309, 1964; *Spectrochim. Acta* 3, 40, 1974; *Z. Analyt. Chem.* 209, 1, 1965; *Pure Appl. Chem.* 34, 35, 1973.

16. A. L. Wilson, *Analyst* 86, 72, 1961.

17. IUPAC, *Spectrochim. Acta* 33 B 242, 1978.

18. M. Otto, *Chemometrics, Statistics and Computer Application in Analytical Chemistry,* Wiley-VCH, Weinheim, 1999.

19. C. Bala, L. Mutihac, *Chemometrie, Principii şi aplicaţii în chimia analitică,* Ed. Universităţii din Bucureşti, 2000.

20. G. Ehrlich, R. Gherbatsch, *Z. Analyt. Chem.* 220, 260, 1966.

21. C. Liteanu, I. Rîcă, *Microchim. Acta* 745, 1973; *Pure Appl. Chem.* 44, 535, 1975; *Microchim. Acta* 311, 1975.

22. C. Liteanu, I. Florea, *Microchim. Acta* 983, 1966.

23. G. L. Long, J. D. Winefordner, *Anal. Chem.* **55**, 712 A, 1983.

https://biblioteca-digitala.ro / https://unibuc.ro

3. Noțiuni de statistică matematică utilizate în chimia analitică

3.1. Generalități

Sistemele analitice reprezintă, conform teoriei sistemelor elaborate de Ludwig von Bertalanffy, sisteme informaționale. Rezultatele finale ale funcționării lor constau din informațiile obținute despre structura şi/sau compoziția calitativă şi cantitativă a materialelor analizate. Ele se obțin prin decodificarea datelor experimentale (spectre, voltamograme, cromatograme) înregistrate la analiza probelor de interes.

Rezultatele analitice primare prezintă variabilitate chiar dacă măsurătorile repetate se efectuează cu atenție deosebită, asupra aceloraşi probe şi în aceleaşi condiții experimentale. Acest fapt nu poate fi evitat şi complică interpretarea datelor. Totuşi, chimistul analist poate extrage informații pertinente din rezultatele unor experiențe afectate de erori prin utilizarea ***statisticii matematice***. Aceasta constituie o ramură a matematicii aplicate la studiul fenomenelor de masă dependente de mai mulți factori. Ea cuprinde metodele de calcul prin care se obțin valorile cele mai probabile ale caracteristicilor variabile ale fenomenelor studiate.

Acest capitol prezintă noțiunile fundamentale de statistică matematică necesare la rezolvarea unui mare număr de probleme analitice (calibrarea instrumentelor de măsură, calculul rezultatelor determinărilor, evaluarea caracteristicilor de performanță ale instrumentelor şi metodelor de analiză, elaborarea noilor metode de analiză, optimizarea proceselor analitice etc).

3.1.1. Populație şi selecție

Prin ***populație statistică*** se înțelege totalitatea indivizilor de acelaşi fel. Termenul de indivizi nu trebuie însă înțeles în accepțiunea sa uzuală. El poate reprezenta, de exemplu, evenimente descrise vag (precum prezența sau absența din probe a unor compuşi) sau în formă cuantificabilă (de exemplu, conținutul în probe a unuia sau a mai multor substanțe chimice). În orice caz, indivizii unei populații statistice trebuie să prezinte trăsături similare, să fie caracterizați de aceleaşi caracteristici. Astfel, dacă aceştia reprezintă materiale de analizat - trebuie să aparțină aceleiaşi tipologii (de exemplu, alimente sau formațiuni geologice de acelaşi tip etc).

https://biblioteca-digitala.ro / https://unibuc.ro

Deoarece valorile caracteristicilor indivizilor dintr-o populație statistică prezintă variabilitate, distribuția lor se descrie probabilistic prin funcții de distribuție (vezi secțiunea 3.2). Acestea nu pot fi însă deduse pe baza unei cunoașteri exhaustive a tuturor indivizilor din populația statistică deoarece populațiile statistice sunt compuse, în general, dintr-un număr foarte mare de indivizi sau, uneori, acest număr este chiar infinit (de exemplu, atunci când populația statistică reprezintă rezultatele posibil de a fi obținute la analiza printr-o metodă chimică specificată a unui material analizat). De aceea, ele se pot estima doar probabilistic, prin analiza unui număr finit de indivizi aleşi aleator din ansamblul populației statistice. Totalitatea acestora reprezintă o ***selecție*** iar numărul lor se numeşte ***volumul selecției***.

3.1.2. Variabile aleatoare

Caracteristicile indivizilor ce alcătuiesc populațiile statistice reprezintă variabile aleatoare. După valorile pe care le iau, acestea pot fi ***discrete*** (de exemplu, numărul dezintegrărilor nucleare ale unui material radioactiv într-o perioadă de timp) sau ***continue***. Matematic, variabilele aleatoare sunt specificate complet prin valorile pe care acestea le iau şi prin probabilitățile asociate diferitelor valori.

Variabilele aleatoare continue se definesc, cel mai frecvent, pe toată scala numerelor reale. Întrucât aceasta conține o mulțime infinită de valori, se pot calcula doar probabilitățile de a găsi valorile variabilei aleatoare în intervale specificate utilizând pentru aceasta funcțiile densitate de probabilitate (vezi secțiunile următoare). Variabilele aleatoare discontinue se specifică prin două şiruri de numere, unul reprezentând valorile variabilei iar celălalt – probabilitățile acestora. Astfel, dacă variabila aleatoare discontinuă X ia valorile x_1, x_2, ..., x_n ale căror probabilități sunt p_1, p_2, ..., p_n ($p_1+p_2+...+p_n=1$), aceasta va fi reprezentată convențional prin:

$$X: \begin{pmatrix} x_1 & x_2 & & x_n \\ p_1 & p_2 & & p_n \end{pmatrix}$$

Valorile pe care le iau cele mai multe variabile aleatoare întâlnite în practică sunt grupate în jurul unei valori medii, m_X. Este uşor de demonstrat că, pentru variabila discretă prezentată mai înainte, media se calculează astfel:

https://biblioteca-digitala.ro / https://unibuc.ro

$$m_X = \sum_{i=1}^{n} p_i x_i$$

Pentru caracterizarea împrăştierii valorilor variabilelor aleatoare în jurul mediei se utilizează ***dispersia*** sau ***varianţa***, acestea reprezentând două sinonime pentru aceeaşi mărime statistică notată σ_X^2. În cazul anterior, dispersia este dată de ecuaţia:

$$\sigma_X^2 = \sum_{i=1}^{n} p_i (x_i - m_X)^2$$

Frecvent, în simbolurile mărimilor ce caracterizează distribuţia variabilelor aleatoare se omite denumirea variabilei. Pentru variabile aleatoare continue, mediile şi dispersiile acestora se pot calcula pornind de la funcţiile de distribuţie a probabilităţii, după cum va rezulta din secţiunile următoare.

3.2. Distribuţii de probabilitate

Dacă se repetă un experiment analitic, rezultatele acestuia alcătuiesc o mulţime de valori distribuite mai mult sau mai puţin întâmplător pe scala numerelor reale. Repartiţia tuturor valorilor care s-ar măsura într-un număr de experienţe foarte mare este descrisă probabilistic prin ***funcţia de distribuţie a probabilităţii***, notată F. În continuare, această funcţie va fi numită, pe scurt, ***funcţie de probabilitate***.

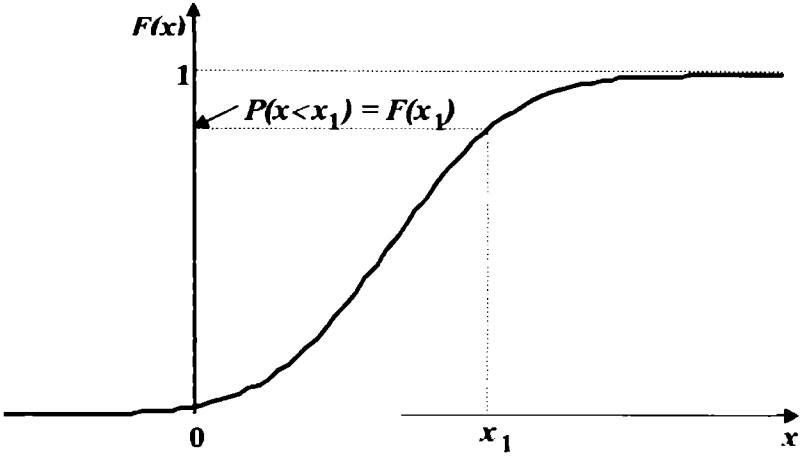


Figura 3.1. Funcţia de distribuţie a unei variabile aleatoare distribuite normal.

https://biblioteca-digitala.ro / https://unibuc.ro

Prin definiție, funcția de probabilitate reprezintă probabilitatea ca variabila aleatoare x să fie mai mică decât o anumită valoare. Aşadar, probabilitatea ca variabila x să fie mai mică decât o valoare x_1 este:

$$P(x < x_1) = F(x_1)$$

Funcția de distribuție este crescătoare şi ia valori cuprinse în domeniul [0,1]. Un exemplu de astfel de funcție este prezentat în figura 3.1.

Prin definiție, derivata funcției de distribuție, F, reprezintă ***funcția densitate de distribuție a probabilității***, p:

$$p(x) = \frac{dF(x)}{dx}$$

Această funcție, care va fi denumită în continuare ***funție densitate de probabilitate***, este mult utilizată în cazul variabilelor aleatoare continue. Astfel, probabilitatea ca o variabilă aleatoare de acest tip să ia o valoare situată în domeniul (x_1, x_2) este dată de integrala:

$$P(x_1 < x < x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} p(x)\,dx$$

, iar media m si dispersia σ^2 se calculează astfel:

$$m = \int_{-\infty}^{+\infty} xp(x)dx = M(x); \qquad \sigma^2 = \int_{-\infty}^{+\infty} [x - M(x)]^2 p(x)dx = D(x)$$

Dispersia variabilei aleatoare x, $D(x)$, se mai numeşte şi momentul centrat de ordinul doi.

Distribuțiile de probabilitate se utilizează curent la prelucrarea rezultatelor experimentale pentru obținerea unor informații de calitate. Aceasta se face prezentă la estimări de parametri, la evaluarea intervalelor de încredere a parametrilor estimați şi la verificări de ipoteze în legătură cu parametrii estimați.

3.2.1. Funcția de distribuție normală

În concepția lui Gauss, funcția de distribuție normală poate fi luată drept model pentru studiul probabilistic al proceselor de măsură. Acest mod de distribuție al varia-

https://biblioteca-digitala.ro / https://unibuc.ro

bilelor aleatoare este fundamental în statistica matematică şi în teoria erorilor şi are următoarea expresie a funcţiei densitate de probabilitate:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty$$

, în care x este variabila aleatoare, m este media distribuţiei (momentul de ordinul unu), iar σ^2 este dispersia (momentul centrat de ordinul doi).

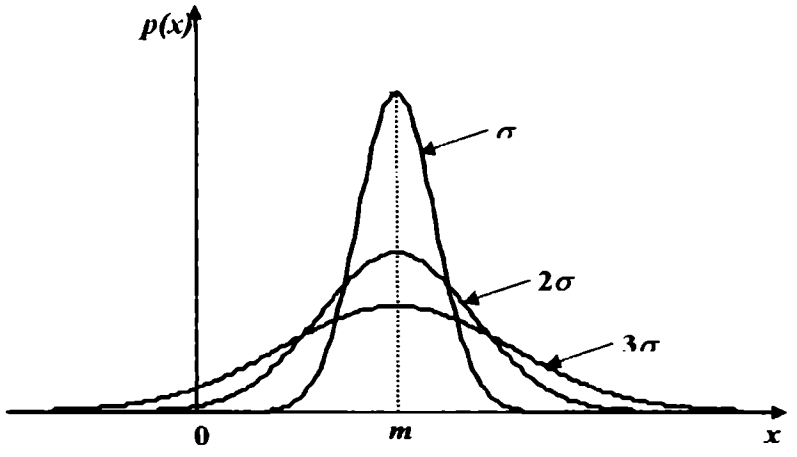


Figura 3.2. Curbele unor funcţii densitate de probabilitate normală cu aceeaşi medie m şi cu dispersii diferite.

Funcţia densitate de probabilitate a unei variabile aleatoare distribuite normal este specificată complet dacă se cunosc media m şi abaterea standard σ ale variabilei aleatoare. În aceste condiţii, se spune ca variabila aleatoare are o distribuţie $N(m, \sigma)$.

Funcţia densitate de probabilitate normală are următoarele proprietăţi mai importante:

- admite un maxim pentru $x = m$,
- este simetrică în raport cu m,
- modificarea parametrului m determină deplasarea curbei normale de-a lungul abscisei fără a-i afecta forma,

https://biblioteca-digitala.ro / https://unibuc.ro

- modificarea lui σ conduce la lățirea sau îngustarea curbei fără ca media m să fie afectată (vezi figura 3.2),

- curba are două puncte de inflexiune corespunzătoare valorilor $x = m \pm \sigma$.

Se disting două teoreme a căror utilitate este importantă pentru multe din aplicațiile distribuției normale:

1. Media a n valori de selecție având distribuția normală $N(m, \sigma)$ are la rândul său o distribuție normală $N(m, \sigma/\sqrt{n})$.

2. Media a n valori de selecție aparținând unei distribuții nenormale, pentru n număr mare, are o distribuție normală sau apropiată de aceasta.

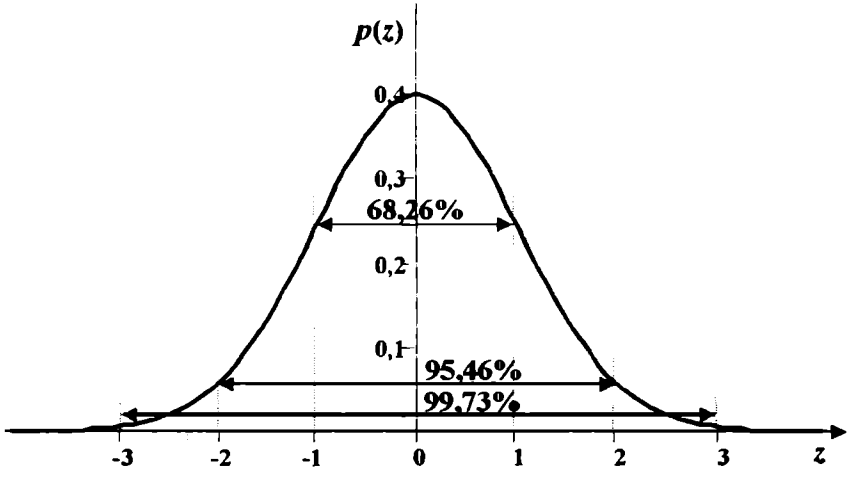


Figura 3.3. Funcția densitate de probabilitate normală standard şi valorile probabilității ca variabila aleatoare z să fie cuprinsă în domeniile (-1, 1), (-2, 2) şi, respectiv, (-3, 3).

Dacă x este o variabilă aleatoare cu funcția de distribuție $N(m, \sigma)$, atunci variabila aleatoare z (abaterea normată) definită prin: $z = \frac{x - m}{\sigma}$

, are o distribuție $N(0, 1)$ cu funcția densitate de probabilitate normală standard (sau normată):

https://biblioteca-digitala.ro / https://unibuc.ro

$$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right], \qquad -\infty < z < \infty$$

pentru care $m = 0$ şi $\sigma^2 = 1$. Această funcţie este prezentată în figura 3.3.

Funcţia de probabilitate a distribuţiei normale standard este:

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{z^2}{2}\right) dz$$

Pentru $z > 0$, $F(z) = 0{,}5 + P(z)$ şi, datorită proprietăţilor de simetrie, $F(-z) = 0{,}5 - P(z)$, unde $P(z) = \frac{1}{\sqrt{2\pi}} \int_{0}^{z} \exp\left(-\frac{z^2}{2}\right) dz$ reprezintă funcţia lui Laplace. Valori ale acestei funcţii sunt prezentate în tabelul 3.1.

Funcţia densitate de probabilitate normată este simetrică în raport cu $z = 0$. Ea a fost acceptată drept model probabilistic pentru evaluarea, prelucrarea şi cercetarea fluctuaţiilor determinărilor analitice deoarece, în condiţiile sistemelor analitice bine concepute şi cu o execuţie corectă a analizelor, rezultatele experimentale se distribuie normal sau apropiat de aceasta. Abaterile pronunţate ale rezultatelor analitice de la o distribuţie normală pune în evidenţă nereguli ale sistemului analitic studiat.

3.2.2. Funcţia de distribuţie χ^2

Dacă $x_1, x_2, \ldots, x_n$ sunt variabile aleatoare independente cu distribuţii normale normate, $N(0, 1)$, variabila aleatoare χ^2 prezentată mai jos:

$$\chi^2 = \sum_{i=1}^{n} x_i^2; \quad 0 < \chi^2 < +\infty$$

, are o distribuţie χ^2 cu $\nu = (n-1)$ grade de libertate şi o funcţie densitate de probabilitate (vezi figura 3.4) dată de ecuaţia:

$$p(\chi^2, \nu) = \frac{(\chi^2)^{(\nu/2)-1} \exp(-\chi^2/2)}{2^{\nu/2}\Gamma(\nu/2)}; \quad \nu > 0; \quad 0 \le \chi^2 < +\infty$$

, unde $\Gamma(z) = \int_{0}^{\infty} t^{z-1} \exp(-t) dt$ reprezintă funcţia Gamma de speţa a doua a lui Euler.

Media acestei distribuţii este egală cu ν iar dispersia - cu 2ν.

https://biblioteca-digitala.ro / https://unibuc.ro

Tabelul 3.1

Valorile funcției Laplace*, $P(z)=\frac{1}{\sqrt{2\pi}}\int_0^z \exp\left(-\frac{z^2}{2}\right)dz$

$P(z)$	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993

* Argumentele, z, reprezintă suma numerelor din prima coloană şi din primul rând, situate pe acelaşi rând şi, respectiv, în aceeaşi coloană cu valorile funcției.

https://biblioteca-digitala.ro / https://unibuc.ro

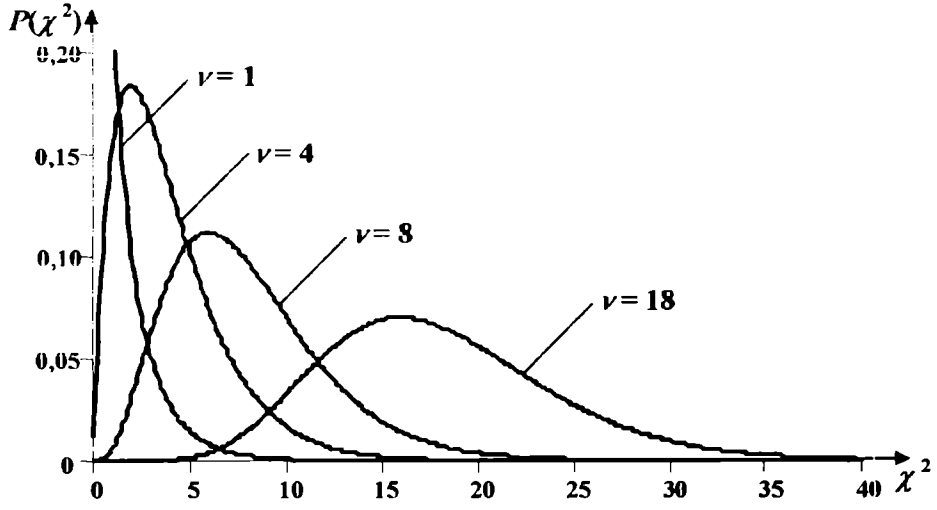


Figura 3.4. Funcția densitate de probabilitate a distribuției χ^2 pentru patru grade de libertate diferite.

Funcția de probabilitate a distribuției χ^2 este:

$$F(\chi^2,\nu)=\int_0^{\chi^2}\frac{(\chi^2)^{(\nu/2)-1}\exp(-\chi^2/2)}{2^{\nu/2}\Gamma(\nu/2)}d\chi^2$$

Pentru zece valori diferite ale acestei funcții, valorile variabilei χ^2 sunt prezentate în tabelul 3.2 pentru diferite grade de libertate ν.

Tabelul 3.2

Valorile variabilei χ^2 pentru $P = F(\chi^2,\nu) = \int_0^{\chi^2} p(\chi^2,\nu)d\chi^2$

$\nu \backslash P$	0,005	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995
1	0,0000393	0,0001571	0,000982	0,00393	0,01579	2,71	3,84	5,02	6,63	7,88
2	0,01003	0,0201	0,0506	0,1026	0,211	4,61	5,99	7,38	9,21	10,60
3	0,0717	0,1148	0,216	0,352	0,584	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,064	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,610	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,237	1,635	2,20	10,64	12,59	14,45	16,81	18,55
7	0,989	1,239	1,690	2,17	2,83	12,02	14,07	16,01	18,48	20,3
8	1,344	1,646	2,18	2,73	3,49	13,36	15,51	17,53	20,1	22,0
9	1,735	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,7	23,6
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,5	23,2	25,2

https://biblioteca-digitala.ro / https://unibuc.ro

Tabelul 3.2

(Continuare)

$\nu \setminus P$	0,005	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995
11	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,9	24,7	26,8
12	3,07	3,57	4,40	5,23	6,30	18,55	21,0	23,3	26,2	28,3
13	3,57	4,11	5,01	5,89	7,04	19,81	22,4	24,7	27,7	29,8
14	4,07	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1	31,3
15	4,60	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6	32,8
16	5,14	5,81	6,91	7,96	9,31	23,5	26,3	28,8	32,0	34,3
17	5.70	6.41	7.56	8.67	10.09	24.8	27.6	30.2	33.4	35.7
18	6,26	7,01	8,23	9,39	10,86	26,0	28,9	31,5	34,8	37,2
19	6,84	7,63	8,91	10,12	11,65	27,2	30,1	32,9	36,2	38,6
20	7,43	8,26	9,59	10,85	12,44	28,4	31,4	34,2	37,6	40,0
22	8,64	9,54	10,98	12,34	14,04	30,8	33,9	36,8	40,3	42,8
24	9,89	10,86	12,40	13,85	15,66	33,2	36,4	39,4	43,0	45,6
26	11,16	12,20	13,84	15,38	17,29	35,6	38,9	41,9	45,6	48,3
28	12,46	13,56	15,31	16,93	18,94	37,9	41,3	44,5	48,3	51,0
30	13,79	14,95	16,79	18,49	20,6	40,3	43,8	47,0	50,9	53,7
35	17,19	18,51	20,6	22,5	24,8	46,1	49,8	53,2	57,3	60,3
40	20,7	22,2	24,4	26,5	29,1	51,8	55,8	59,3	63,7	66,8
45	24,3	25,9	28,4	30,6	33,4	57,5	61,7	65,4	70,0	73,2
50	28,0	29,7	32,4	34,8	37,7	63,2	67,5	71,4	76,2	79,5
55	31,7	33,6	36,4	39,0	42,1	68,8	73,3	77,4	82,3	85,7
60	35,5	37,5	40,5	43,2	46,5	74,4	79,1	83,3	88,4	92,0
65	39,4	41,4	44,6	47,4	50,9	80,0	84,8	89,2	94,4	98,1
70	43,3	45,4	48,8	51,7	55,3	85,5	90,5	95,0	100,4	104,2
75	47,2	49,5	52,9	56,1	59,8	91,1	96,2	100,8	106,4	110,3
80	51,2	53,5	57,2	60,4	64,3	96,6	101,9	106,6	112,3	116,3
85	55,2	57,6	61,4	64,7	68,8	102,1	107,5	112,4	118,2	122,3
90	59,2	61,8	65,6	69,1	73,3	107,6	113,1	118,1	124,1	128,3
95	63,2	65,9	69,9	73,5	77,8	113,0	118,8	123,9	130,0	134,2
100	67,3	70,1	74,2	77,9	82,4	118,5	124,3	129,6	135,8	140,2
110	75,6	78,5	82,9	86,8	91,5	129,4	135,5	140,9	147,4	151,9
120	83.9	86.9	91.6	95.7	100.6	140.2	146.6	152.2	159.0	163.6
130	92,2	95,5	100,3	104,7	109,8	151,0	157,6	163,5	170,4	175,3
140	100,7	104,0	109,1	113,7	119,0	161,8	168,6	174,6	181,8	186,8
150	109,1	112,7	118,0	122,7	128,3	172,6	179,6	185,8	193,2	198,4
160	117,7	121,3	126,9	131,8	137,5	183,3	190,5	196,9	205	210
170	126,3	130,1	135,8	140,8	146,8	194,0	201	208	216	221
180	134,9	138,8	144,7	150,0	156,2	205	212	219	227	233
190	143,5	147,6	153,7	159,1	165,5	215	223	230	238	244
200	152,2	156,4	162,7	168,3	174,8	226	234	241	249	255

https://biblioteca-digitala.ro / https://unibuc.ro

Funcția de distribuție χ^2 poate fi utilizată pentru descrierea densităților de probabilitate a dispersiilor de selecție s^2, ceea ce permite evaluarea unor intervale de încredere pentru dispersii și, respectiv, pentru abaterile standard. De asemenea, cu ajutorul distribuției χ^2 se poate construi un test de semnificație pentru compararea a două dispersii de selecție.

Pentru utilizarea distribuției χ^2 la verificarea ipotezelor referitoare la funcțiile de distribuție ale variabilelor aleatoare, se iau în considerație următoarele aspecte. Fiind dată o variabilă aleatoare discretă x cu o distribuție:

$$\begin{pmatrix} x_1 & x_2 & & x_m \\ p_1 & p_2 & & p_m \end{pmatrix}$$

, asupra căreia se fac n experiențe independente, variabila x luând de n_1 ori valoarea x_1, de n_2 ori valoarea x_2, . . . , de n_m ori valoarea x_m, astfel încât $n_1 + n_2 + ... + n_m = n$, se poate arăta că, dacă $n \to \infty$, variabila:

$$\frac{(n_1 - np_1)^2}{np_1} + \frac{(n_2 - np_2)^2}{np_2} + ... + \frac{(n_m - np_m)^2}{np_m}$$

, tinde către o repartiție χ^2, cu parametrii $\nu = n - 1$ grade de libertate.

3.2.3. Funcția de distribuție *t* (Student)

Variabila aleatoare t a funcției de distribuție Student reprezintă raportul a două variabile aleatoare: z - variabila normală standard cu o funcție de distribuție normală N(0, 1) și, respectiv, variabila aleatoare $\sqrt{\chi^2(\nu)/\nu}$, independentă de prima:

$$t = \frac{z}{\sqrt{\chi^2/\nu}} = \frac{z}{s/\sigma} = \frac{\bar{x} - m}{\sigma}\frac{\sigma}{s} = \frac{\bar{x} - m}{s}$$

Funcția densitate de probabilitate pentru variabila t are expresia următoare:

$$p(t,\nu) = \frac{1}{\sqrt{\pi\nu}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad -\infty < t < +\infty$$

, unde ν reprezintă numărul gradelor de libertate asociat cu s^2. Aspectul acesteia, pentru trei grade de libertate diferite, poate fi observat în figura 3.5. Când numărul gradelor de libertate $\nu \to \infty$, distribuția t se apropie de o distribuție normală standard.

https://biblioteca-digitala.ro / https://unibuc.ro

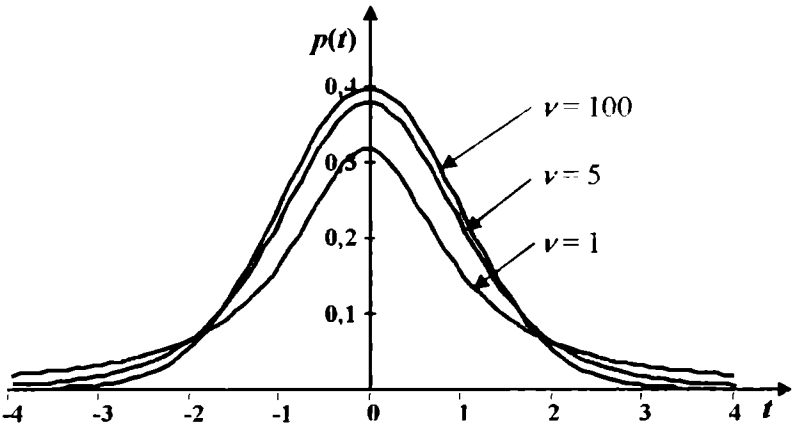


Figura 3.5. Densitățile de probabilitate pentru trei variabile Student.

Funcția de probabilitate a distribuției Student este:

$$F(t,\nu)=\int_{-\infty}^{t} p(t,\nu)dt$$

Această funcție are proprietăți similare cu distribuția normală standard: pentru $t > 0$, $F(t,\nu) = 0{,}5 + P(t,\nu)$ și $F(-t) = 0{,}5 - P(t,\nu)$, unde $P(t,\nu)=\int_{0}^{t} p(t,\nu)dt$.

În tabelul 3.3 sunt tabelate valorile variabilei t pentru diferite grade de libertate și pentru opt valori diferite ale funcției de probabilitate.

În general, funcția de distribuție t se folosește pentru evaluarea intervalului de încredere pentru media unei distribuții normale și pentru verificarea ipotezelor statistice pe care le implică analiza chimică pentru selecții mici (n<30).

3.2.4. Funcția de distribuție *F*

O variabilă aleatoare, F, are o funcție de distribuție F cu ν_1, ν_2 grade de libertate și o funcție de densitate de probabilitate cu următoarea expresie:

$$p(F)=\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\frac{\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{\nu_2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}F^{(\nu_1/2-1)}\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}, \qquad 0<F<\infty$$

https://biblioteca-digitala.ro / https://unibuc.ro

Tabelul 3.3

Valorile variabilei t (Student) pentru $P = F(t, \nu) = \int_{-\infty}^{t} p(t, \nu)dt$

$\nu \setminus P$	0,700	0,800	0,900	0,950	0,990	0,995	0,998	0,999
1	1,963	3,078	6,314	12,706	63,657	127,321	318,31	636,62
2	1,386	1,886	2,920	4,303	9,925	14,089	22,327	31,599
3	1,250	1,638	2,353	3,182	5,841	7,453	10,215	12,924
4	1,190	1,533	2,132	2,776	4,604	5,598	7,173	8,610
5	1,156	1,476	2,015	2,571	4,032	4,773	5,893	6,869
6	1,134	1,440	1,943	2,447	3,707	4,317	5,208	5,959
7	1,119	1,415	1,895	2,365	3,499	4,029	4,785	5,408
8	1,108	1,397	1,860	2,306	3,355	3,833	4,501	5,041
9	1,100	1,383	1,833	2,262	3,250	3,690	4,297	4,781
10	1,093	1,372	1,812	2,228	3,169	3,581	4,144	4,587
11	1,088	1,363	1,796	2,201	3,106	3,497	4,025	4,437
12	1,083	1,356	1,782	2,179	3,055	3,428	3,930	4,318
13	1,079	1,350	1,771	2,160	3,012	3,372	3,852	4,221
14	1,076	1,345	1,761	2,145	2,977	3,326	3,787	4,140
15	1,074	1,341	1,753	2,131	2,947	3,286	3,733	4,073
16	1,071	1,337	1,746	2,120	2,921	3,252	3,686	4,015
17	1,069	1,333	1,740	2,110	2,898	3,222	3,646	3,965
18	1,067	1,330	1,734	2,101	2,878	3,197	3,610	3,922
19	1,066	1,328	1,729	2,093	2,861	3,174	3,579	3,883
20	1,064	1,325	1,725	2,086	2,845	3,153	3,552	3,850
25	1,058	1,316	1,708	2,060	2,787	3,078	3,450	3,725
30	1,055	1,310	1,697	2,042	2,750	3,030	3,385	3,646
35	1,052	1,306	1,690	2,030	2,724	2,996	3,340	3,591
40	1,050	1,303	1,684	2,021	2,704	2,971	3,307	3,551
45	1,049	1,301	1,679	2,014	2,690	2,952	3,281	3,520
50	1,047	1,299	1,676	2,009	2,678	2,937	3,261	3,496
60	1,045	1,296	1,671	2,000	2,660	2,915	3,232	3,460
70	1,044	1,294	1,667	1,994	2,648	2,899	3,211	3,435
80	1,043	1,292	1,664	1,990	2,639	2,887	3,195	3,416
90	1,042	1,291	1,662	1,987	2,632	2,878	3,183	3,402
100	1,042	1,290	1,660	1,984	2,626	2,871	3,174	3,390
110	1,041	1,289	1,659	1,982	2,621	2,865	3,166	3,381
120	1,041	1,289	1,658	1,980	2,617	2,860	3,160	3,373

Funcția de distribuție F se utilizează la verificarea ipotezelor de semnificație a două dispersii de selecție. Dacă se consideră selecțiile $x_{11}, x_{12}, ..., x_{1n}$ și $x_{21}, x_{22}, ..., x_{2n}$ efectuate asupra aceleeași distribuții normale cu aceeași dispersie, variabila:

$$F = s_1^2 / s_2^2 = \left(\chi_1^2 / \nu_1\right) / \left(\chi_2^2 / \nu_2\right)$$

https://biblioteca-digitala.ro / https://unibuc.ro

, are o distribuție F cu $\nu_1 = n_1 - 1$ și $\nu_2 = n_2 - 1$ grade de libertate. Unele valori ale variabilei F sunt prezentate în tabelele 3.4 și 3.5.

Tabelul 3.4

Valorile variabilei F pentru $P = P(F) = \int_0^F p(F;\nu_1,\nu_2)dF = 0{,}95$; $(\nu_1 = 1 \div 10)$

	1	2	3	4	5	6	7	8	9	10
1	161	199	216	225	230	234	237	239	241	242
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16

https://biblioteca-digitala.ro / https://unibuc.ro

Tabelul 3.5

Valorile variabilei F pentru $P = P(F) = \int_0^F p(F;\nu_1,\nu_2)dF = 0{,}95$; ($\nu_1 = 1 \div 10$)

	11	12	13	14	15	16	17	18	19	20
1	243	244	245	245	246	246	247	247	248	248
2	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4
3	8,76	8,74	8,73	8,71	8,70	8,69	8,68	8,67	8,67	8,66
4	5,94	5,91	5,89	5,87	5,86	5,84	5,83	5,82	5,81	5,80
5	4,70	4,68	4,66	4,64	4,62	4,60	4,59	4,58	4,57	4,56
6	4,03	4,00	3,98	3,96	3,94	3,92	3,91	3,90	3,88	3,87
7	3,60	3,57	3,55	3,53	3,51	3,49	3,48	3,47	3,46	3,44
8	3,31	3,28	3,26	3,24	3,22	3,20	3,19	3,17	3,16	3,15
9	3,10	3,07	3,05	3,03	3,01	2,99	2,97	2,96	2,95	2,94
10	2,94	2,91	2,89	2,86	2,85	2,83	2,81	2,80	2,79	2,77
11	2,82	2,79	2,76	2,74	2,72	2,70	2,69	2,67	2,66	2,65
12	2,72	2,69	2,66	2,64	2,62	2,60	2,58	2,57	2,56	2,54
13	2,63	2,60	2,58	2,55	2,53	2,51	2,50	2,48	2,47	2,46
14	2,57	2,53	2,51	2,48	2,46	2,44	2,43	2,41	2,40	2,39
15	2,51	2,48	2,45	2,42	2,40	2,38	2,37	2,35	2,34	2,33
16	2,46	2,42	2,40	2,37	2,35	2,33	2,32	2,30	2,29	2,28
17	2,41	2,38	2,35	2,33	2,31	2,29	2,27	2,26	2,24	2,23
18	2,37	2,34	2,31	2,29	2,27	2,25	2,23	2,22	2,20	2,19
19	2,34	2,31	2,28	2,26	2,23	2,21	2,20	2,18	2,17	2,16
20	2,31	2,28	2,25	2,22	2,20	2,18	2,17	2,15	2,14	2,12
21	2,28	2,25	2,22	2,20	2,18	2,16	2,14	2,12	2,11	2,10
22	2,26	2,23	2,20	2,17	2,15	2,13	2,11	2,10	2,08	2,07
23	2,24	2,20	2,18	2,15	2,13	2,11	2,09	2,08	2,06	2,05
24	2,22	2,18	2,15	2,13	2,11	2,09	2,07	2,05	2,04	2,03
25	2,20	2,16	2,14	2,11	2,09	2,07	2,05	2,04	2,02	2,01
26	2,18	2,15	2,12	2,09	2,07	2,05	2,03	2,02	2,00	1,99
27	2,17	2,13	2,10	2,08	2,06	2,04	2,02	2,00	1,99	1,97
28	2,15	2,12	2,09	2,06	2,04	2,02	2,00	1,99	1,97	1,96
29	2,14	2,10	2,08	2,05	2,03	2,01	1,99	1,97	1,96	1,94
30	2,13	2,09	2,06	2,04	2,01	1,99	1,98	1,96	1,95	1,93

3.2.5. Distribuţia binomială

Funcţia de distribuţie binomială se poate defini ca probabilitatea de realizare de x ori a unui eveniment în n experienţe independente, atunci când probabilitatea de apariţie a evenimentului într-un singur experiment este p. Aceasta are expresia:

$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \qquad \text{unde } x = 0, 1, \ldots, n$$

Pentru $0{,}3 \le p \le 0{,}7$ şi $n \ge 30$, distribuţia binomială este exprimată bine de către o distribuţie normală cu $m = n\,p$ şi dispersia $\sigma^2 = n\,p\,(1\text{-}p)$.

https://biblioteca-digitala.ro / https://unibuc.ro

3.3. Estimarea parametrilor

Estimatorul este o regulă de calcul a unor parametri în baza unei selecții iar ***estimat*** – valoarea care se obține într-un anumit caz.

Un exemplu tipic de estimare îl constituie estimația mediei şi dispersiei distribuției unei variabile aleatoare x cu ajutorul unei selecții x_1, x_2, ..., x_n conform relațiilor de mai jos:

$$x = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

În cazul proceselor analitice sunt numeroase exemple de estimare, cum ar fi: estimarea funcțiilor de calibrare; estimarea parametrilor distribuției rezultatelor şi erorilor de analiză; estimarea performanțelor proceselor analitice; estimarea funcției de răspuns a procesului analitic etc.

Problema estimării parametrilor se prezintă pentru cazul în care între variabilele de interes $\boldsymbol{x}^T = (x_1, x_2, ..., x_m)$ şi răspunsul y există o dependență de tip liniar:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m + e$$

Pentru estimarea parametrilor $\boldsymbol{b}^T = (b_1, b_2, ..., b_m)$, în cazul în care b_0=0, se utilizează o mulțime de perechi de n perechi de date $\boldsymbol{x}(i) = (x_1(i), x_2(i), ..., x_m(i)); y(i)$, care satisfac relația (a se vedea şi capitolul următor):

$$y(i) = \boldsymbol{x}(\boldsymbol{i})\,\boldsymbol{b} + e(i)$$

, unde $e(i)$ este zgomotul cu o valoare de aşteptare egală cu zero.

Se notează cu $\hat{\boldsymbol{b}}$ valorile de estimare pentru parametri iar regula de estimare prin: $\hat{\boldsymbol{b}} = \boldsymbol{G}(\boldsymbol{x}; \boldsymbol{y})$.

Printre cei mai importanți estimatori se pot enumera: estimatori Bayes, estimatorul verosimilității maxime, estimatorul Markov şi estimatorul celor mai mici pătrate.

3.4. Intervale de încredere

Ca noțiune, ***intervalul statistic de încredere*** a fost utilizat pentru prima dată de către Laplace, pentru a mări gradul de semnificație al estimării parametrilor distribuției binomiale. Apoi, Neumann a fost cel care a dezvoltat teoria şi terminologia inter-

https://biblioteca-digitala.ro / https://unibuc.ro

valului de încredere. Noțiunea de ***estimare*** implică noțiunea de ***interval de încredere***. Deci, problema aceasta apare ori de câte ori există o problemă de estimare.

Dacă se consideră o variabilă aleatoare cu o funcție densitate de probabilitate ipotetică $p(x, b)$, parametrul b va fi estimat în baza unei selecții $x(1)$, $x(2)$, ..., $x(n)$ cu ajutorul unui estimator:

$$\hat{b} = G[x(1), x(2), ..., x(n)]$$

Valoarea de estimație $\hat{b}$ este la rândul său o variabilă aleatoare a cărei funcție densitate de probailitate o notăm cu $p(\hat{b}; b)$. Probabilitatea P ca variabila $\hat{b}$ să se găsească într-un interval $b_1 < \hat{b} < b_2$ este dată de integrala:

$$P = P(b_1 < \hat{b} < b_2) = \int_{b_1}^{b_2} p(\hat{b}, b) d\hat{b}$$

Mărimea P este numită şi coeficient de încredere. În general, în statistică problema se pune invers, adică: dacă se dă o probabilitate P, să se stabilească un interval de încredere pentru parametrul existent.

3.4.1. Intervalul de încredere pentru media unei distribuții

Se pot diferenția două situații:

a) *Dispersia distribuției normale este cunoscută.* În această situație, pentru estimarea mediei $\bar{x}$ se efectuează o selecție $x(1)$, $x(2)$, ..., $x(n)$ referitoare la variabila x distribuită $N(m, \sigma)$. La rândul său, media va avea o distribuție $N(m, s/\sqrt{n})$. Variabila normală standard $z = (\bar{x} - m)\sqrt{n}/s$ are o distribuție normală normată $N(0, 1)$. Pentru o probabilitate de încredere, P, mediei m a distribuției îi corespunde un interval de încredere simetric conform expresiei $\bar{x} \pm z_{1-\alpha/2} s/\sqrt{n}$, în care $\alpha = 1 - P$ iar $z_{1-\alpha/2}$ se deduce din condiția:

$$P[z_{1-\alpha/2}] = (1 + P)/2$$

b) *Dispersia distribuției normale este necunoscută.* În acest caz, pentru evaluarea intervalului de încredere simetric corespunzător unei probabilități de încredere $P=1-\alpha$, în cazul unei distribuții $N(m, \sigma)$ căreia nu i se cunoaşte nici media, nici dis-

https://biblioteca-digitala.ro / https://unibuc.ro

persia, se ia în considerare variabila $t = (x - m)/(s/\sqrt{n})$ cu o distribuție Student cu ν=n-1 grade de libertate. În consecință, expresia intervalului de încredere este:

$$\bar{x} \pm t_{1-\alpha/2;\nu} s/\sqrt{n}$$

, unde α=1-P iar $t_{1-\alpha/2;\nu}$ se deduce din ecuația $P[t_{1-\alpha/2;\nu}] = (1+P)/2$.

Acest caz al evaluării intervalului de încredere pentru media unei distribuții normale cu dispersie necunoscută este mult utilizat în chimia analitică.

În același mod se poate calcula intervalul de încredere pentru diferența a două medii, m_1-m_2, corespunzătoare variabilelor aleatoare x și y cu distribuțiile $N(m_1, \sigma)$ și $N(m_2, \sigma)$. Ca aplicație în chimia analitică, acest caz se utilizează în special la evaluarea intervalului de încredere pentru diferența dintre mediile rezultatelor obținute în două laboratoare diferite, prin aceeași metodă.

3.4.2. Intervalul de încredere pentru dispersia de selecție

Pentru estimarea dispersiei unei variabile aleatoare cu o distribuție $N(m, \sigma)$ se consideră o selecție $x(1)$, $x(2)$, ..., $x(n)$ din care se evaluează dispersia de selecție s^2. Se cunoaște faptul că variabila: $\chi^2 = (n-1)s^2/\sigma^2$ are o distribuție χ^2 cu ν=n-1 grade de libertate. Intervalul de încredere pentru dispersie, pentru probabilitatea P=1-α, are expresia:

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2;\nu}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2;\nu}}$$

Dacă se utilizează volume de selecție mari ($n \geq 30$) se demonstrează că abaterea standard de selecție s are o distribuție $N(s; s/\sqrt{n})$. Deci, se poate admite că variabila:

$$z = \frac{s - \sigma}{s/\sqrt{2n}}$$

are o distribuție $N(0, 1)$. În concluzie, pentru o probabilitate de încredere corespunzătoare, P=1-α, se poate evalua un interval de încredere pentru abaterea standard a unei distribuții normale, pentru volume de selecție mari, cu expresia:

$$s \pm z_{1-\alpha/2} s/\sqrt{2\nu}$$

https://biblioteca-digitala.ro / https://unibuc.ro

3.4.3. Intervalul de încredere pentru parametrii distribuției binomiale

În urma unor determinări repetate de n ori, fenomenul care ne interesează s-a manifestat de m ori. Probabilitatea p de manifestare a evenimentului este estimată de raportul $\hat{p} = m/n$. Ținând seama de proprietățile distribuției binomiale, variabila:

$$z = \frac{\hat{p} - p}{[\hat{p}(1-\hat{p})/n]^{1/2}}$$

are aproximativ distribuția $N(0,1)$ și deci intervalul de încredere pentru parametrul p, corespunzător unei probabilități de încredere $P=1-\alpha$, are următoarea expresie:

$$\hat{p} \pm z_{1-\alpha/2}[\hat{p}(1-\hat{p})/n]^{1/2}$$

3.5. Testarea ipotezelor

În general, testările de ipoteze se aplică la rezolvarea problemelor de evaluare și optimizare a proceselor analitice. Orice decizie luată în privința optimizării unui proces analitic trebuie elaborată în baza unor teste statistice. Teoria testelor de verificare a ipotezelor a fost elaborată de Neumann și Pearson și se aplică mult în analiza chimică.

În chimia analitică, cazul general este reprezentat de testarea unei ipoteze $\boldsymbol{H_0}$ în raport cu alte ipoteze alternative: $\boldsymbol{H_1}$, $\boldsymbol{H_2}$, ... În cazul unei metode noi de analiză, se pune de obicei problema testării ipotezei $\boldsymbol{H_0}$ – media erorilor de analiză este egală cu zero (lipsesc erorile sistematice) în raport cu ipotezele alternative: $\boldsymbol{H_1}$ – media erorilor de analiză < 0 și, respectiv, $\boldsymbol{H_2}$ – media erorilor de analiză > 0. În altă situație analitică, se pune problema testării ipotezei $\boldsymbol{H_0}$ – introducerea sau modificarea unui factor în procesul de analiză duce la îmbunătățirea performanțelor acestuia.

Pentru a studia ipoteza $\boldsymbol{H_0}$ se consideră că parametrul b al procesului luat în studiu are valoarea $b=b_0$ în raport cu ipoteza $\boldsymbol{H_1}$ pentru care $b \neq 0$. Evident, pentru rezolvarea acestei probleme, se efectuează o selecție în baza căreia se obține o valoare $\hat{b}$ pentru parametrul b. Variabilei $\hat{b}$ îi corespunde, conform teoriei statistice, densitatea de probabilitate $p(\hat{b})$. Întrebarea este cât de mare trebuie să fie diferența $\hat{b}$ -b_0 astfel ca ipoteza $\boldsymbol{H_0}$ să fie respinsă? În figura 3.6 este prezentată densitatea de proba-

https://biblioteca-digitala.ro / https://unibuc.ro

bilitate $p(\hat{b})$ pentru cazul când ipoteza H_0 este adevărată. Probabilitatea ca estimatul b să ia o valoare mai mică decât un prag $b_{\alpha/2}$ este:

$$P(b \le b_{\alpha/2}) = \int_{-\infty}^{b_{\alpha/2}} p(\hat{b})d\hat{b} = \alpha/2$$

iar probabilitatea ca estimatul b să ia o valoare mai mare decât un prag $b_{1-\alpha/2}$ este:

$$P(b \ge b_{\alpha/2}) = \int_{b_{1-\alpha/2}}^{+\infty} p(\hat{b})d\hat{b} = \alpha/2$$

În concluzie, pentru formularea deciziilor se selectează o valoare denumită nivelul de semnificație al testului așa încât probabilitatea de a depăși pragurile $b_{\alpha/2}$ și $b_{1-\alpha/2}$ să aibă o valoare prestabilită egală cu 0,10, 0,05 sau 0,01. Dacă valoarea de estimare $\hat{b}$ este mai mică decât $b_{\alpha/2}$ sau mai mare decât $b_{1-\alpha/2}$, ipoteza H_0 este respinsă, iar dacă se situează în regiunea $b_{\alpha/2} < \hat{b} < b_{1-\alpha/2}$, ipoteza este acceptată.

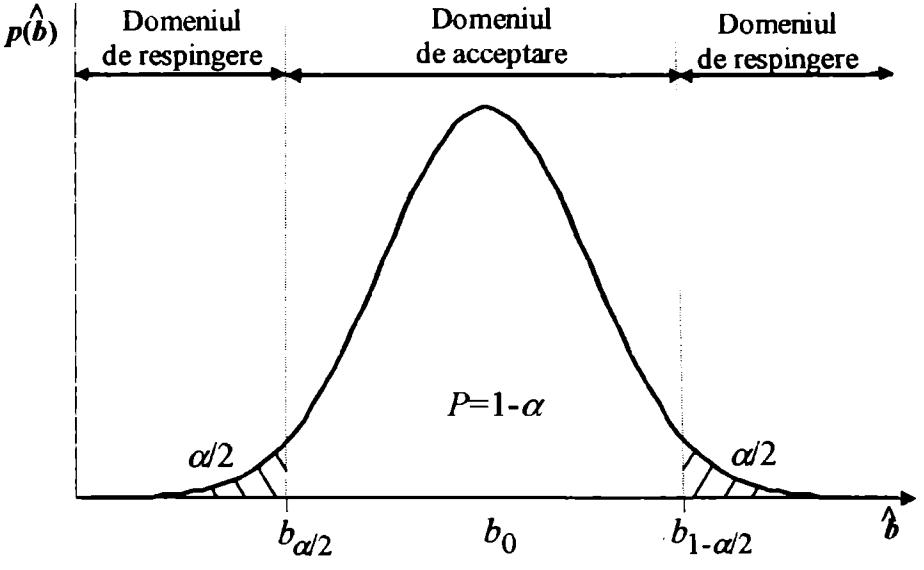


Figura 3.6. Reprezentarea schematică a procesului de testare a ipotezelor.

https://biblioteca-digitala.ro / https://unibuc.ro

Evident, pe baza rezultatelor testului, ipoteza propusă este acceptată sau respinsă. Respingerea ipotezei este o consecință a faptului că rezultatele diferă semnificativ. Esențial este faptul că testele nu sunt folosite pentru demonstrarea unei ipoteze, ci pentru îndrumarea unor acțiuni. În general, la testarea ipotezelor statistice pot să apară două tipuri de erori: ***eroare de tip I***, când se respinge o ipoteză adevărată, și ***eroare de tip II***, când se admite o ipoteză falsă.

Pentru verificarea ipotezelor, lucrările de statistică prezintă mai multe criterii (teste). În funcție de problemele abordate în procesul analitic, se selectează testul potrivit. De exemplu, pentru eliminarea valorilor extreme dintr-un șir de rezultate se folosesc următoarele criterii: criteriul t (Student), criteriul R (Grubbs), criteriul Q (Dixon) etc (vezi Liteanu, *Teoria și metodologia statistică a analizei urmelor* – 1979).

Pentru verificări de ipoteze în legătură cu media unei distribuții sau cu mediile a două distribuții, se ia frecvent ca model distribuția normală (vezi și secțiunile anterioare). Aceste teste utilizează de obicei selecții cu un volum bine precizat. Dacă însă volumul selecției este variabil, se utilizează teste secvențiale sau teste neparametrice elaborate special pentru distribuții specifice. Pentru testarea a mai mult de două medii se utilizează metoda analizei de dispersie.

3.6. Analiza de dispersie în chimia analitică

Analiza de dispersie a fost introdusă de Fischer. Cu toate că este o metodă eficientă de evaluare și de optimizare a proceselor analitice, cu aplicații numeroase în biologie și chimie, studiile referitoare la acest subiect sunt relativ puține. ***Analiza de dispersie*** este o tehnică statistică utilizată la evaluarea unor factori suspecți de a distorsiona veridicitatea rezultatelor. În cazul chimiei analitice, analiza de dispersie este utilizată la identificarea și evaluarea celor mai semnificative surse de erori. Din punct de vedere al metodologiei, se poate menționa analiza de dispersie monofactorială, bifactorială, trifactorială etc.

Studiile menționate în literatura de specialitate a analizei de dispersie prezintă exemple de utilizare a analizei de dispersie monofactorială și bifactorială privind descompunerea dispersiei rezultatelor de analiză în două componente (de exemplu: între zile și, respectiv, în cursul aceleiași zile de lucru) sau în trei componente. În acest

https://biblioteca-digitala.ro / https://unibuc.ro

sens, analiza de dispersie se utilizează foarte mult în studii interlaboratoare. Totuşi, aplicaţiile sale depăşesc cu mult acest domeniu, ea putând fi utilizată inclusiv în probleme de validare a modelelor de regresie. Un exemplu de acest tip este descris în capitolul următor.

Bibliografie

1. M. A. Sharaf, D. L. Illman şi B. R. Kowalski, ***Chemometrics***, Wiley, New York, 1986.

2. N. E. Nahi, ***Estimation Theory and Applications***, Wiley, New York, 1969.

3. P. Eykhoff, ***Identificarea sistemelor, estimarea parametrilor şi stărilor pentru sistemele tehnice, economice, biologice***, Editura Tehnică, Bucureşti, 1977. Traducere din limba engleză.

4. F. R. Hirsch, *Anal. Chem.*, **49**, 691A, 1977.

5. J. H. Riddick, R. Flora şi Q. L. Van Meter, *Clin. Chem.*, **18**, 250, 1972.

6. G. Gootschalk, *Z. Anal. Chem.*, **276**, 81, 1975.

7. C. Liteanu şi I. Rîcă, ***Teoria şi metodologia statistică a analizei urmelor***, Editura Scrisul Românesc, Craiova, 1979.

8. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte şi L. Kaufman, ***Chemometrics: A. Textbook***, Elsevier, Amsterdam, 1988.

9. R. Shiavi, ***Introduction to Applied Statistical Analysis***, Aksen Publishers, Boston, 1991.

10. C. Liteanu şi I. Rîcă, ***Optimizarea proceselor analitice***, Editura Academiei Române, Bucureşti, 1985.

11. D. Ceauşescu, ***Utilizarea statisticii matematice în chimia analitică***, Editura Tehnică, Bucureşti, 1982.

12. M. Otto, ***Chemometrics, Statistics and Computer Application in Analytical Chemistry***, Wiley – VCH, Weinheim, 1998.

https://biblioteca-digitala.ro / https://unibuc.ro

4. Modelarea datelor analitice prin regresii mono sau multivariabile

Un instrument matematic utilizat frecvent în aplicațiile științifice îl reprezintă analiza de regresie. Pentru aceasta, în limba engleză se utilizează, cu valoare aproape sinonimă, mai multe expresii, precum ***empirical modeling***, ***curve fitting***, ***regression analysis*** sau ***forecasting***.

Acest capitol prezintă noțiunile fundamentale ce conduc la regresii de calitate, ele putând avea aplicații la evaluarea funcțiilor de etalonare a sistemelor analitice, la evidențierea și corectarea fenomenelor de interferență, la stabilirea condițiilor optime de lucru pentru sistemele analitice de măsură și la evaluarea performanțelor calitative ale sistemelor analitice. În general, ele pot fi utile ori de câte ori este nevoie să fie modelate matematic seturi de date experimentale.

În secțiunea finală, se discută succint direcțiile actuale de estimare simultană a concentrațiilor mai multor analiți pe baza unui mare număr de date experimentale achiziționate asupra probelor.

4.1. Tipuri de regresii

Răspunsul sistemelor analitice (potențial de electrod, absorbanță, aria picurilor cromatografice, rezultatul unei metode de analiză etc), notat prescurtat y, este influențat, în general, de un număr foarte mare de factori experimentali (concentrații de analiți și de interferenți, condiții de operare și de mediu ș.a.m.d.). De obicei, se dorește modelarea răspunsului în funcție de un număr limitat de factori în condițiile în care toți ceilalți sunt menținuți constanți. Pentru aceasta, factorii experimentali considerați la elaborarea modelului (***variabilele independente***) sunt variați sistematic și se determină valoarea răspunsului (***variabila dependentă***). Între perechile de date astfel obținute există o relație de forma:

$$y_i = f(x_{1i}, x_{2i}, ..., x_{mi}; b_0, b_1, ..., b_m) + e_i; \quad i = 1, n \tag{4.1}$$

unde y_i reprezintă valoarea răspunsului obținută în experimentul numărul i; x_{ji} – valorile din experimentul numărul i a factorilor x_j ($j = 1,m$); b_k – constante denumite parametrii modelului ($k = 0,m$); $f(x_{ji}; b_k)$ – componenta deterministă a valorii răspunsu-

https://biblioteca-digitala.ro / https://unibuc.ro

lui şi e_i – componenta stochastică sau aleatoare. Această ultimă componentă modelează variabilitatea datelor analitice care se manifestă ca urmare a fluctuaţiilor condiţiilor experimentale (temperatură, umiditate, tensiune de alimentare a instrumentelor de măsură etc) pe parcursul determinărilor.

Analiza de regresie constă în modelarea componentei deterministe a semnalelor în prezenţa componentei stochastice sau aleatoare iar tipul de regresie care trebuie utilizat într-un anumit caz depinde de problema de rezolvat. Acesta se referă la numărul variabilelor dependente şi independente din expresia matematică a modelului, precum şi la natura relaţiei de dependenţă dintre variabila independentă şi parametrii b_k. Principalele tipuri de regresii întâlnite în analiza chimică sunt prezentate sintetic în tabelul 4.1.

Tabelul 4.1
Principalele tipuri de regresii

Nr. crt.	Tip de regresie	Caracteristici	Forma modelului
1.	Monovariabilă liniară	Variabile independente: 1 Parametri: 1 sau 2	$y = b_0 + b_1 x$
2.	Monovariabilă polinomială	Variabile independente: 1 Parametri: 2 sau mai mulţi, m+1 (m–gradul polinomului)	$y = b_0 + \sum_{k=1}^{m} b_k x^k$
3.	Monovariabilă neliniară	Variabile independente: 1 Parametri: 2 sau mai mulţi	$y = f(x)$ f – funcţii neliniare
4.	Multivariabilă liniară	Variabile independente: 2 Parametri: 2 sau mai mulţi, m+1 (m–gradul polinomului)	$y = b_0 + \sum_{i=1}^{m} b_i x_i$ Frecvent $b_0 = 0$

Regresia după o singură variabilă (monovariabilă) presupune existenţa unei singure variabile independente, indiferent care este relaţia de dependenţă dintre aceasta şi variabila dependentă, în timp ce în regresia de mai multe variabile (multivariabilă) se presupune că variabila dependentă este o funcţie de mai multe variabile independente. În continuare, peste tot se va nota cu y variabila dependentă în timp ce pentru

https://biblioteca-digitala.ro / https://unibuc.ro

variabilele independente vom folosi notațiile x în regresia monovariabilă și, respectiv, x_1, x_2, x_3, ..., x_m în regresia multivariabilă. În cazul repetării experimentelor, valorile concrete luate de către variabilele independentă și dependente vor fi reprezentate printr-un indice adițional notat, în caz general, cu litera l. De asemenea, acolo unde este necesar, cu z vom reprezenta variabila dependentă transformată iar cu u – variabila independentă transformată.

După forma matematică a modelului utilizat, putem face distincție între modele de regresie ***liniare*** și ***neliniare***. În cazul regresiei liniare, variabila dependentă poate fi reprezentată printr-o combinație liniară de operatori care depind, nu neapărat liniar, de variabila independentă (sau de variabilele independente). În cazul regresiei monovariabile, astfel de operatori pot fi x^c (c – constantă), $\log x$, $1/x$, $\sin x$ etc.

O clasă importantă de modele sunt cele întrebuințate în regresia polinomială monovariabilă, care se aplică frecvent pentru a aproxima, pe anumite porțiuni, funcții neliniare. În acest scop, se pot utiliza și funcțiile spline de ordinul trei.

În cazul în care modelul matematic polinomial se rezumă la primii doi termeni, se obține tipul de regresie liniară propriu zisă. În acest caz, datele experimentale se înscriu pe o dreaptă. Acest ultim tip de regresie este foarte mult utilizat în studii de calibrare ale metodelor de analiză, când variabila independentă semnifică concentrația componentului analizat din etaloane iar b_0 și b_1 – ordonata la origine a dreptei de calibrare și, respectiv, panta acesteea, ultima reprezentând o caracteristică foarte importantă a unei metode de analiză, legată direct de sensibilitatea metodei.

Regresia multivariabilă se aplică frecvent în cazul metodelor spectrometrice de determinare a mai multor componenți din soluții (determinări multicomponent) când se presupune de regulă că $b_0 = 0$.

Metodele neliniare de regresie utilizează pentru exprimarea relației de dependență dintre variabila dependentă și variabila independentă ecuații care nu sunt liniare în toți parametrii. Totuși, ele pot fi liniarizate prin dezvoltări în serie. Astfel, o funcție oarecare depinzând de mai mulți parametri, b_1, b_2, ..., b_m, și având forma:

$$y = f(x; b_1, ..., b_m)$$

https://biblioteca-digitala.ro / https://unibuc.ro

, poate fi bine aproximată, în jurul unui punct din spațiul parametrilor, $\boldsymbol{b_0}$, definit prin vectorul linie: $\boldsymbol{b_0}^T = (b_{10}, b_{20}, \ldots, b_{m0})$, printr-o dezvoltare în serie Taylor din care se reţin termenii de ordinul 0 şi 1, conform ecuaţiei:

$$y = f(x; b_1, \ldots, b_m) + \left(\frac{\delta f}{\delta b_1}\right)_{b=b_0} \cdot \Delta b_1 + \left(\frac{\delta f}{\delta b_2}\right)_{b=b_0} \cdot \Delta b_2 + \ldots + \left(\frac{\delta f}{\delta b_m}\right)_{b=b_0} \cdot \Delta b_m \quad (4.2)$$

, cu condiţia ca Δb_k ($k = 1,m$) să fie suficient de mici. Utilizând această proprietate, regresia neliniară poate fi efectuată în mod similar celei liniare în cazul în care iniţial se cunoaşte sau se poate determina o estimaţie suficient de bună a valorilor parametrilor modelului.

Ca exemple de modele monoparametrice neliniare, aplicate frevcent în chimia analitică, putem menţiona:

- funcţii exponenţiale:

$$y = b_1 \exp(b_2 x)$$

, care descriu funcţiile de răspuns ale instrumentelor analitice după aplicarea unui semnal în formă de treaptă;

- funcţii de tip Gauss:

$$y = b_1 \exp\{-[b_2(x - b_3)]^2\}$$

, care au aplicaţii numeroase în statistica matematică (vezi capitolul precedent) şi descriu bine multe semnale cu aspect de pic, inclusiv pe cele cromatografice;

- funcţii de tip Lorenz:

$$y = \frac{b_1}{1 + b_2(x - b_3)^2}$$

, care descriu forma picurilor de rezonanţă magnetică nucleară;

- funcţii putere:

$$y = b_1 x^{b_2}$$

, care descriu intensitatea liniilor spectrale în spectrografia atomică de emisie;

- funcţii logaritmice:

$$y = b_1 + b_2 \log x$$

, care descriu răspunsul electrozilor ion-selectivi.

https://biblioteca-digitala.ro / https://unibuc.ro

Unele ecuații neliniare pot fi aduse ușor, prin transformări de variabile, la o formă liniară. Astfel, în cazul potențiometriei, unde relația de dependență potențial – concentrație, dată de ecuația lui Nernst, este logaritmică, regresia liniară se poate aplica după ce se efectuează următoarea transformare a variabilei independente:

$$u = \log x$$

4.2. Etapele analizei de regresie

Având determinat un set de date, analiza de regresie presupune parcurgerea următoarelor etape: *definirea modelului*, *estimarea parametrilor* (regresia propriu zisă), *validarea modelului* și *calculul intervalelor de încredere* ale parametrilor.

4.2.1. Definirea modelului

Această etapă este foarte importantă deoarece stabilește tipul de regresie care va fi utilizat la estimarea parametrilor. Ea presupune explicitarea relației de dependență dintre variabilele dependentă și, respectiv, independente (estimarea funcției *f*) și, implicit, stabilirea numărului de parametri ai modelului (constantele b_k).

Modelul matematic este uneori sugerat de legile fizice care se respectă în cazul investigat (de exemplu, în cazul calibrărilor, legea Lambert-Beer în spectrometrie, ecuația lui Nernst în potențiometrie etc). Alteori, modelul cel mai probabil se impune cu necesitate. Astfel, dacă se dorește identificarea tipurilor de interferențe care afectează o metodă de analiză, se compară rezultatele obținute cu metoda studiată pentru un set de probe, y_i, cu rezultatele obținute, pentru aceleași probe, cu o metodă de referință considerată exactă, x_i. Un model liniar de forma:

$$y = b_0 + b_1 x$$

este probabil să se respectate în astfel de cazuri. Dacă modelul este validat, valorile estimate ale parametrilor furnizează următoarele concluzii:

a) Dacă $b_0 = 0$ și $b_1 = 1$, metoda este exactă;

b) Dacă $b_0 \neq 0$ și $b_1 = 1$, metoda este afectată de o eroare sistematică de tip aditiv și, pentru obținerea rezultatului corect, b_0 trebuie scăzut din rezultatul calculat;

c) Dacă $b_0 = 0$ și $b_1 \neq 1$, metoda este afectată de o eroare de tip multiplicativ și, pentru obținerea rezultatului corect, se poate aplica metoda adițiilor standard;

https://biblioteca-digitala.ro / https://unibuc.ro

d) Dacă $b_0 \neq 0$ şi $b_1 \neq 1$, metoda este afectată atât de erori de tip aditiv, cât şi multiplicativ.

Există şi situaţii, întâlnite mai ales la optimizarea parametrilor de lucru ai metodelor de analiză, când relaţia dintre răspuns şi factorii de optimizat nu este cunoscută a priori şi când se recurge la modelare empirică. Inspectarea vizuală a graficelor reprezentând variaţia răspunsului în funcţie de factorii consideraţi sugerează funcţiile matematice care vor fi utilizate la definirea modelului în astfel de cazuri.

La aplicarea regresiei monovariabile liniare se presupune că între perechile de date experimentale, (y_1, x_1), (y_2, x_2), ... , (y_n, x_n), măsurate în condiţii reale şi, prin urmare, afectate de zgomot, există următoarele relaţii:

$$y_1 = b_0 + b_1 x_1 + e_1$$
$$y_2 = b_0 + b_1 x_2 + e_2$$
$$\ldots\ldots\ldots\ldots$$
$$y_i = b_0 + b_1 x_i + e_i$$
$$\ldots\ldots\ldots\ldots$$
$$y_n = b_0 + b_1 x_n + e_n$$

Ecuaţiile de mai sus pot fi scrise în formă condensată utilizând notaţii matriceale. Astfel, notând cu $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{e}$ şi $\boldsymbol{X}$ vectorii coloană ai măsurătorilor, ai parametrilor, ai erorilor (sau ai reziduurilor) şi, respectiv, matricea variabilelor independente, definiţi prin ecuaţiile următoare:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ .. \\ y_i \\ .. \\ y_n \end{pmatrix}; \quad \boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}; \quad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ .. \\ e_i \\ .. \\ e_n \end{pmatrix}; \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & .. \\ 1 & x_i \\ 1 & .. \\ 1 & x_n \end{pmatrix}$$

, sistemul de ecuaţii de mai sus este echivalent cu:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \qquad (4.3)$$

https://biblioteca-digitala.ro / https://unibuc.ro

Peste tot în acest capitol se adoptă convenţiile uzuale de scriere a mărimilor vectoriale şi matriceale. Astfel, literele mici scrise îngroşat reprezintă vectori coloană iar majusculele – mărimi matriceale. Transpusa vectorilor sau matricelor se reprezintă adăugând litera $\boldsymbol{T}$ în dreapta, sus la simbolul iniţial. De exemplu, dacă $\boldsymbol{y}$ reprezintă primul vector coloană cu n elemente definit puţin mai sus, atunci $\boldsymbol{y}^T$ reprezintă următorul vector linie:

$$\boldsymbol{y}^T = \begin{pmatrix} y_1 & y_2 & \cdots & y_i & \cdots & y_n \end{pmatrix}$$

De asemenea:

$$\boldsymbol{X}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_i & \cdots & x_n \end{pmatrix}$$

Estimaţiile oricăror mărimi, indiferent de natura acestora, se vor reprezenta prin adăugarea unui accent circumflex (^) deasupra simbolului mărimii respective. Astfel, estimaţiile obţinute pe baza modelului de regresie şi adoptat pentru vectorii măsurătorilor şi ai parametrilor se vor nota cu $\hat{\boldsymbol{y}}$ şi, respectiv, cu $\hat{\boldsymbol{b}}$.

Şi pentru celelalte tipuri de regresii prezentate în tabelul 4.1 se pot scrie ecuaţii de dependenţă între variabilele dependente măsurate în prezenţa zgomotului şi cele independente asemănătoare ca formă cu ecuaţiile din regresia monovariabilă liniară. Singura deosebire constă în modificarea componentelor deterministe care derivă din forma modelelor de regresie utilizate.

Pentru ca ecuaţiile astfel obţinute să depindă liniar de toţi parametrii şi în cazul regresiei neliniare, expresia modelului utilizat în acest caz se poate înlocui cu aproximarea sa dată de ecuaţia 4.2 dacă sunt îndeplinite condiţiile de valabilitate a acesteia menţionate anterior.

Deşi relaţiile dintre valorile măsurătorilor efectuate şi a factorilor experimentali consideraţi diferă de la un tip de regresie la altul, totuşi, ecuaţia 4.3 este generală. Ea se aplică în toate cazurile de regresie liniară şi, în condiţiile restrictive menţionate, la regresia neliniară cu condiţia să se utilizeze expresii adecvate pentru vectorul parametrilor şi pentru matricea măsurătorilor. Forma matematică a acestora este prezentată în tabelul 4.2 pentru tipurile de regresie prezentate în tabelul 4.1 (altele decât regresia

https://biblioteca-digitala.ro / https://unibuc.ro

monovariabilă liniară deja discutată) presupunând că modelul matematic de regresie utilizat este cel din ultima coloană a tabelului 4.1.

Tabelul 4.2
Forma vectorului parametrilor şi a matricei măsurătorilor în trei tipuri diferite de regresii

Nr. crt.	Tip de regresie	Forma vectorului $\boldsymbol{b}$	Forma matricei $\boldsymbol{X}$
1.	Monovariabilă polinomială	$\begin{pmatrix} b_0 \\ b_1 \\ . \\ b_k \\ \\ b_m \end{pmatrix}$	$\begin{pmatrix} 1 & x_1 & x_1^j & x_1^m \\ . & & & \\ 1 & x_i & x_i^j & x_i^m \\ . & & . & . \\ 1 & x_n & x_n^j & x_n^m \end{pmatrix}$
2.	Monovariabilă neliniară	$\begin{pmatrix} 1 \\ \Delta b_1 \\ . \\ \Delta b_k \\ . \\ \Delta b_m \end{pmatrix}$	$\begin{pmatrix} f_1 & \frac{\delta f_1}{\delta b_1} & \frac{\delta f_1}{\delta b_k} & \frac{\delta f_1}{\delta b_m} \\ . & & & \\ f_i & \frac{\delta f_i}{\delta b_1} & \frac{\delta f_i}{\delta b_k} & \frac{\delta f_i}{\delta b_m} \\ . & & & \\ f_n & \frac{\delta f_n}{\delta b_1} & \frac{\delta f_n}{\delta b_k} & \frac{\delta f_n}{\delta b_m} \end{pmatrix}$
3.	Multivariabilă liniară	$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \\ . \\ b_m \end{pmatrix}$	$\begin{pmatrix} 1 & x_{11} & x_{21} & x_{m1} \\ 1 & x_{12} & x_{22} & x_{m2} \\ 1 & x_{13} & x_{23} & x_{m3} \\ . & & & \\ . & & & \\ 1 & x_{1n} & x_{2n} & x_{mn} \end{pmatrix}$

În cazul regresiei monovariabile neliniare, toate elementele matricei $\boldsymbol{X}$ (f_i şi $\delta f_i/\delta b_k$) reprezintă modelul de regresie, f, sau derivatele de ordinul unu ale acestuia în raport cu diferiţi parametri şi se calculează la valorile x_i ale variabilei independente. De asemenea, toate acestea se calculează în punctul $\boldsymbol{b_0}$, acesta reprezentând vectorul estimaţie iniţială a parametrilor modelului.

https://biblioteca-digitala.ro / https://unibuc.ro

4.2.2. Estimarea parametrilor modelului

În această etapă se calculează valorile necunoscute ale parametrilor prin suprapunerea cât mai bună a modelului peste perechile de date ceea ce presupune minimizarea unui criteriu matematic a cărui expresie depinde de natura distribuției erorilor în setul de date.

Cel mai frecvent, se consideră că valorile factorilor, x_{ji}, sunt cunoscute fără eroare, acestea necontribuind la diferențele dintre valorile măsurate ale răspunsului, y_i, și cele estimate pe baza modelului, $\hat{y}_i$, diferențe denumite frevent deviații sau reziduuri. În majoritatea aplicațiilor din chimia analitică, unde x_j reprezintă concentrații de analiți, această presupunere este adevărată deoarece de regulă precizia răspunsului analitic este considerabil mai redusă decât cea obținută la prepararea soluțiilor standard.

Dacă modelul adoptat aproximează suficient de bine dependența exprimată în relația 4.2 prin funcția f pe tot domeniul de variație al factorilor, deviațiile $(y_i - \hat{y}_i)$ reprezintă o estimație a erorilor experimentale (e_i). Criteriul matematic, $\boldsymbol{R}$, prin minimizarea căruia se estimează corect valorile parametrilor modelului îl reprezintă suma pătratelor deviațiilor:

$$\boldsymbol{R} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

cu condiția ca erorile experimentale de determinare a răspunsului să reprezinte variabile aleatoare care să îndeplinească următoarele condiții:

a) să fie distribuie normal cu media zero;

b) să fie independente între ele;

c) să aibă aceeași varianță, σ^2.

Suma pătratelor deviațiilor, $\boldsymbol{R}$, reprezintă norma vectorilor reziduurilor, notată $\|\boldsymbol{e}\|$ și calculată, în notații matriceale, cu expresia:

$$\|\boldsymbol{e}\| = \boldsymbol{e}^T \boldsymbol{e} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T (\boldsymbol{y} - \hat{\boldsymbol{y}}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \qquad (4.4)$$

Aceasta este o funcție dependentă de parametrii modelului, b_k. Valorile cele mai probabile ale acestora, reprezentate prin următorul vector: $\hat{\boldsymbol{b}}^T = (\hat{b}_1, \hat{b}_2, ..., \hat{b}_m)$, se estimează rezolvând următorul sistem de ecuații normale liniare:

https://biblioteca-digitala.ro / https://unibuc.ro

$$\partial \boldsymbol{R} / \partial b_k = 0, \qquad k = 0, m \tag{4.5}$$

, care reprezintă condiția de minim a criteriului $\boldsymbol{R}$.

Numărul de ecuații din sistemul precedent este egal cu numărul de parametri ai modelului, m+1. În toate cazurile de regresie liniară, soluția generală a acestor ecuații este dată de expresia:

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{4.6}$$

, în care $\boldsymbol{X}^T \boldsymbol{X}$ reprezintă inversa generalizată a matricei $\boldsymbol{X}$ de tip Moore-Penrose. Este preferabil ca această matrice să aibă o formă cât mai simplă, ceea ce se poate realiza adoptând un plan experimental corespunzător. În cazul calibrărilor unor determinări multicomponent, aceasta presupune prepararea unor standarde conținând, rând pe rând, câte un singur component.

Indiferent de numărul liniilor şi coloanelor matricei $\boldsymbol{X}$, care nu trebuie să fie neapărat egale, matricea $\boldsymbol{X}^T \boldsymbol{X}$ are o formă pătrată şi, pentru ca soluția reprezentată prin ecuațiile 4.6 să existe, trebuie să fie inversabilă (determinant diferit de zero).

În cazul regresiei neliniare, estimarea parametrilor urmează o cale întrucâtva diferită de cea prezentată anterior. Dacă în ecuația 4.4 se utilizează vectorul $\hat{\boldsymbol{y}}$ exprimat prin modelul de regresie neliniar adoptat, sistemul de ecuații 4.5 nu mai are o formă liniară uşor de rezolvat. Presupunând însă că se poate determina, indiferent prin ce mijloace, o estimație inițială suficient de bună a parametrilor modelului neliniar, notată $\boldsymbol{b}_0$, $\hat{\boldsymbol{y}}$ poate fi înlocuit în ecuația 4.4 prin $\boldsymbol{X}\boldsymbol{b}$, unde $\boldsymbol{X}$ este matricea măsurătorilor cu expresia dată în tabelul 4.2 pentru regresia neliniară iar $\boldsymbol{b}$ este definit prin relația:

$$\boldsymbol{b}^T = (1, \boldsymbol{b}_0^T) = (1, b_{10}, b_{20}, \ldots, b_{m0})$$

În aceste condiții, se verifică uşor că ecuația 4.4 devine:

$$\|\boldsymbol{e}\| = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = (\boldsymbol{y} - \hat{\boldsymbol{y}} - \boldsymbol{J}\Delta\boldsymbol{b})^T (\boldsymbol{y} - \hat{\boldsymbol{y}} - \boldsymbol{J}\Delta\boldsymbol{b}) = (\Delta\boldsymbol{y} - \boldsymbol{J}\Delta\boldsymbol{b})^T (\Delta\boldsymbol{y} - \boldsymbol{J}\Delta\boldsymbol{b}) \tag{4.7}$$

unde: $\Delta\boldsymbol{y}$ reprezintă vectorul abaterilor valorilor măsurate ale răspunsului, $\boldsymbol{y}$, față de valorile estimate pe baza modelului neliniar adoptat, $\hat{\boldsymbol{y}}$, în care s-au utilizat valorile estimate inițial ale parametrilor; $\Delta\boldsymbol{b}$ – vectorul diferențelor dintre valorile adevărate ale parametrilor şi cele estimate inițial iar $\boldsymbol{J}$ este matricea lui Jacobi definită prin:

https://biblioteca-digitala.ro / https://unibuc.ro

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial b_1} & \frac{\partial f_1}{\partial b_k} & \frac{\partial f_1}{\partial b_m} \\ \frac{\partial f_i}{\partial b_1} & \frac{\partial f_i}{\partial b_k} & \frac{\partial f_i}{\partial b_m} \\ \frac{\partial f_n}{\partial b_1} & \frac{\partial f_n}{\partial b_k} & \frac{\partial f_n}{\partial b_m} \end{pmatrix}$$

Elementele acestei matrici au aceleaşi semnificaţii ca şi elementele corespunzătoare din matricea X corespunzatoare (vezi secţiunea 4.3.1 şi tabelul 4.2).

Cu aceste observaţii, sistemul de ecuaţii normale 4.5, în care criteriul R se înlocuieşte cu ecuaţia 4.7, se aplică şi în regresia neliniară, în acelaşi mod ca şi în cea liniară. Rezolvarea sa conduce la vectorul $\Delta\hat{b}$. Prin similitudine cu ecuaţia 4.6, expresia sa este dată de:

$$\Delta\hat{b} = (J^T J)^{-1} J^T \Delta y$$

Vectorul $\hat{b} = b_0 + \Delta\hat{b}$ reprezintă de obicei o estimaţie a parametrilor modelului neliniar mai bună decât cea iniţială, b_0, dacă aceasta din urmă este suficient de exactă. Pentru a obţine o estimaţie şi mai bună a parametrilor modelului, algoritmul de regresie neliniară poate fi reluat prin înlocuirea estimaţiei iniţiale a parametrilor, b_0, cu ultima estimaţie obţinută, $\hat{b}$. Un număr mic de iteraţii (3 – 5) conduce, de regulă, la valorile adevărate ale parametrilor modelului. Aceste valori se recunosc uşor deoarece două iteraţii succesive conduc practic la acelaşi rezultat. Totuşi, în cazul în care estimaţia iniţială a parametrilor, b_0, diferă mult de valorile reale, algoritmul de regresie menţionat poate să eşueze, fapt care se recunoaşte prin lipsa de convergenţă a valorilor estimate ale parametrilor obţinute prin iteraţii succesive.

Datorită formei criteriului de minimizat, R, metodele de estimare a parametrilor prezentate se numesc ***regresii prin metoda celor mai mici pătrate***.

În timp ce proprietăţile a) şi b) ale erorilor experimentale menţionate anterior sunt îndeplinite în practică în cvasitotalitatea cazurilor, ultima proprietate, denumită ***homoscedasticitate***, nu este de obicei respectată într-un domeniu larg de variaţie al factorilor. De aceea, se recomandă testarea statistică a acesteea ori de câte ori există

https://biblioteca-digitala.ro / https://unibuc.ro

dubii asupra valabilității sale. Pentru aceasta, pentru fiecare pereche de valori a factorilor, $x_i^T = (x_{1i}, \quad x_{2i}, \qquad x_{mi})$, se repetă experimentul analitic de mai multe ori măsurându-se valorile răspunsului, y_{ii}, de fiecare dată, apoi se calculează dispersiile și, în final, se testează egalitatea acestora (de exemplu, cu ajutorul testului lui Bartlett).

În cazul multor tehnici analitice, incluzând măsurătorile radiometrice și spectrometrice, varianța măsurătorilor experimentale este dependentă de valoarea răspunsului, y_i, proprietatea aceasta numindu-se ***heteroscedasticitate***. Aceasta se întâmplă datorită tipurilor specifice de zgomot care afectează măsurătorile. Astfel, în măsurători radiometrice, bazate pe numărarea particulelor care ajung la detector, predomină un tip de zgomot supus distribuției lui Poisson. În astfel de cazuri, dispersia erorilor experimentale este proporțională cu rădăcina pătrată a valorilor semnalelor chiar în prezența altor tipuri de erori (în afara celor de numărare datorate detectorului) cum ar fi: erori de pipetare, variații ale timpilor de reacție sau a temperaturii etc. În tehnicile spectrometrice, cum ar fi absorbția atomică, zgomotul datorat fluctuațiilor intensității sursei de lumină este proporțional cu intensitatatea semnalelor. În alte tehnici analitice poate să predomine zgomotul datorat componentelor electronice ale instrumentului de măsură care poate avea varianță constantă sau dependentă de valoarea semnalelor.

În cazul heteroscedasticității nu se mai poate aplica metoda celor mai mici pătrate și, pentru efectuarea regresiei, se apelează la una din următoarele alternative: aplicarea metodei celor mai mici pătrate ponderate sau transformarea variabilelor pentru obținerea homoscedasticității.

La aplicarea metodei celor mai mici pătrate ponderate trebuie cunoscute dispersiile, s_i, pentru valorile răspunsului, notat y_i, în fiecare punct, x_i, din domeniul factorilor. Pentru aceasta este nevoie de un număr, notat n_i, suficient de mare de valori, y_{ii}, ale răspunsului, măsurate prin repetarea experimentului la fiecare set de valori ale factorilor, acesta constituind un important neajuns al metodei.

Trebuie menționat că dacă numerele de repetări, n_i, sunt prea mici, la testarea egalității mediilor aplicând testul Bartlett, probabilitatea respingerii ipotezei heteroscedasticității, în cazul în care aceasta este adevărată, este destul de mare. Totuși, dacă dispersiile, s_i, s-au estimat suficient de corect, criteriul de optimizat, $\boldsymbol{R}$, are expresia:

https://biblioteca-digitala.ro / https://unibuc.ro

$$\boldsymbol{R} = \sum_{i=1}^{n} (w_i e_i)^2 = \sum_{i=1}^{n} [w_i (\bar{y}_i - \hat{y}_i)]^2 \qquad (4.8)$$

, unde w_i reprezintă factorii de pondere ai reziduurilor, daţi de relaţia:

$$w_i = \frac{1}{s_i}$$

, $\bar{y}_i$ – media măsurătorilor individuale ale variabilei dependente obţinute la aceleaşi valori ale variabilelor independente:

$$\bar{y}_i = \sum_{l=1}^{n_i} y_{il}$$

, iar restul mărimilor au semnificaţii cunoscute.

Cu următoarele notaţii:

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & w_n \end{pmatrix}; \qquad \boldsymbol{U} = \boldsymbol{W}^T \boldsymbol{W} = \begin{pmatrix} w_1^2 & 0 & 0 & 0 \\ 0 & w_2^2 & 0 & 0 \\ 0 & 0 & . & 0 \\ 0 & 0 & 0 & w_n^2 \end{pmatrix}$$

, soluţia sistemului de ecuaţii normale 4.5, în care s-a înlocuit criteriul $\boldsymbol{R}$ prin expresia 4.8, are forma:

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T \boldsymbol{U} \boldsymbol{X})^{-1} \left(\boldsymbol{X}^T \boldsymbol{U}\, \boldsymbol{y}\right)$$

Ideea de bază a ***regresiei ponderate prin metoda celor mai mici pătrate*** este, aşadar, de a atribui cea mai mare importanţă datelor măsurate cu cea mai mare precizie. Cea mai dificilă problemă de rezolvat la aplicarea acestui tip de regresii constă în estimarea corectă a factorilor de pondere deoarece este nevoie de un număr mare de repetări ale experimentelor, ceea ce poate fi costisitor şi/sau poate necesita mult timp.

În absenţa unui volum de date suficient de mare, se poate presupune, luând în considerare şi experienţa anterioară desprinsă din tipuri similare de experienţe, o relaţie funcţională între varianţă şi răspuns. Astfel, se cunoaşte că în radioimunoanaliză s_i depinde aproape liniar de $\sqrt{y_i}$. Un alt caz foarte important este cel în care s_i depinde liniar de y_i şi în care deviaţia relativă standard este constantă, aşa cum se întâmplă la concentraţii ridicate, în analiza prin ICP. În astfel de cazuri se poate aplica o transformare convenabilă a variabilei dependente în vederea obţinerii homoscedasticităţii.

https://biblioteca-digitala.ro / https://unibuc.ro

Dacă $s_i \propto \sqrt{y_i}$, homoscedasticitatea se obține prin transformarea $z = \sqrt{y}$, iar dacă $s_i \propto y_i$, se utilizează în același scop transformarea $z = \log y$.

Transformarea variabilei dependente modifică foarte mult forma modelului inițial. Uneori este convenabil să se aplice o transformare și variabilei independente, x. Astfel, o transformare dublu logaritmică ($z = \log y$, $u = \log x$) în cazul regresiei monovariabile, transformă o dreaptă care trece prin origine (așa cum este cazul în majoritatea calibrărilor) într-o altă dreaptă de forma: $z = a + u$.

După aplicarea unei transformări asupra variabilei dependente care conduce la homoscedasticitate, parametrii noului model obținut se estimează prin regresia celor mai mici pătrate aplicate noii variabile, z.

Pentru ilustrarea proprietăților de homo- și heteroscedasticitate, în figura 4.1 se prezintă trei cazuri diferite de dependență a dispersiei de valoarea răspunsului în cazul regresiei liniare monovariabile.

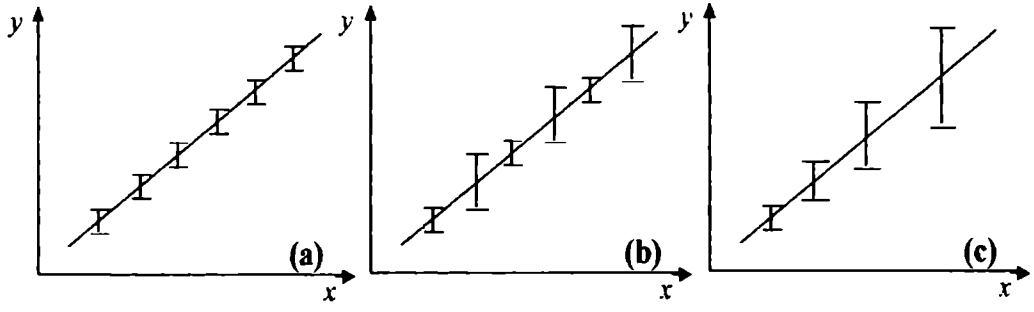


Figura 4.1. Ilustrarea proprietăților de homo- și heteroscedasticitate. a) Homoscedasticitate; b) Heteroscedasticitate; c) Heteroscedasticitate cu deviație relativă standard constantă.

Revenind la exemplul din secțiunea 4.3.1, în care se discută identificarea tipurilor de erori sistematice care afectează metodele noi de analiză, se constată că nu sunt îndeplinite condițiile de aplicare a metodelor de regresie prezentate anterior, respectiv presupunerea că variabila independentă nu este afectată de erori este falsă. Într-adevăr,

https://biblioteca-digitala.ro / https://unibuc.ro

chiar dacă metoda de referință este exactă, ceea ce implică lipsa erorilor sistematice, rezultatele obținute cu ajutorul său, x_i, sunt afectate de erori întâmplătoare care nu pot fi neglijate. În cazul în care metoda nouă și cea de referință prezintă precizii foarte asemănătoare, estimarea parametrilor se va realiza prin minimizarea sumei pătratelor distanțelor $|M_iO_i|$ dintre punctele experimentale și model calculate în direcție perpendiculară pe dreapta de regresie, așa cum se sugerează în figura 4.2. Acest caz este mult mai dificil de tratat matematic comparativ cu cele discutate anterior și nu va fi aprofundat în acest capitol.

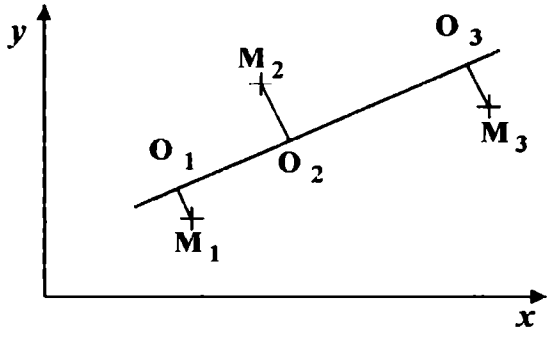


Figura 4.2. Modul de efectuare a regresiei când variabila independentă, x, este afectată de erori comparabile cu ale variabilei dependente, y.

4.2.3. Validarea modelului

Această etapă are ca scop testarea concordanței dintre modelul matematic obținut în etapa anterioară și perechile de date experimentale. Dacă nu se identifică o lipsă de concordanță între acestea, presupunerile referitoare la model sunt corecte și acesta descrie bine relația deterministă existentă între răspuns și factorii experimentali considerați. Doar astfel de modele vor fi utilizate ulterior pentru calculul concentrațiilor de analit în probe noi sau pentru stabilirea optimului de funcționare al unor sisteme analitice, în cazul în care modelul a fost elaborat într-un experiment de calibrare sau, respectiv, de optimizare.

https://biblioteca-digitala.ro / https://unibuc.ro

Pentru validarea modelului se pot utiliza metode grafice sau statistice. Acestea pornesc de la premiza că, dacă modelul este corect, deviațiile dintre datele experimentale și cele calculate pe baza modelului sunt variabile aleatoare cu caracteristici apropiate de a erorilor experimentale care afectează măsurarea valorilor răspunsului. Evident, metodele statistice sunt de preferat, acestea fiind mai obiective. Dar indiferent ce metode se aplică în această etapă, trebuie știut că acestea sunt foarte insensibile când numărul perechilor de date este redus.

După calcularea reziduurilor se poate testa natura distribuției acestora, fie grafic, prin reprezentarea frecvențelor de apariție a acestora, fie statistic, aplicând testul χ^2. În cazul în care ipoteza distribuției normale a reziduurilor nu poate fi respinsă, modelul poate fi validat.

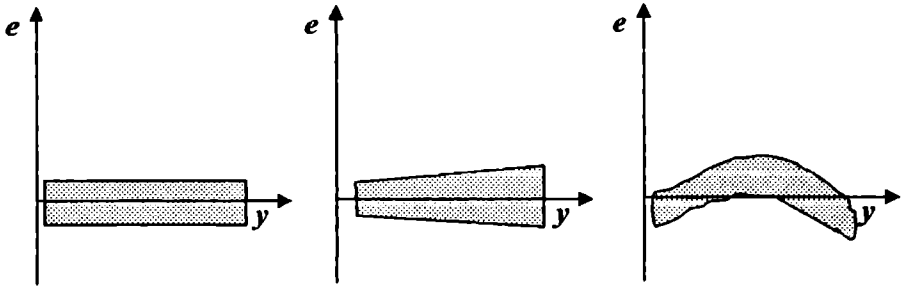


Figura 4.3. Exemple de forme de distribuție a reziduurilor.

Reziduurile calculate mai pot fi reprezentate grafic în funcție de valorile estimate, sau măsurate experimental, ale răspunsului (y_i sau $\hat{y}_i$), sau de valorile factorilor (de exemplu, x_i – în cazul regresiei monovariabile) și se analizează forma benzii în care acestea sunt dispersate. Următoarele situații, dintre care primele trei sunt sugerate în figura 4.3, sunt mai probabile:

a) Reziduurile se distribuie aleator într-o bandă cu media zero și cu lărgime constantă, situație care poate apărea în caz de homoscedasticitate și dacă modelul este corect;

https://biblioteca-digitala.ro / https://unibuc.ro

b) Reziduurile se distribuie aleator într-o bandă cu media zero şi cu lărgime variabilă, această situaţie putând descrie un caz de heteroscedasticitate, dacă modelul este corect;

c) Reziduurile se distribuie aleator într-o bandă a cărei medie, în unele porţiuni, se depărtează în mod sistematic de la valoarea zero, caz în care modelul va fi invalidat;

d) Majoritatea reziduurilor sunt distribuite într-o bandă de forma celor din figura 4.3 (a) şi (b), cu excepţia unora care se situează mult în afara limitelor benzii. Perechile de date, (y_i, x_{ij}), care conduc la valori foarte mari ale reziduurilor în comparaţie cu majoritatea celorlalte, pot fi considerate valori aberante. Acestea se pot elimina din setul de date iniţial şi apoi se recalculează, mai exact, valorile parametrilor. De asemenea, intervalele de încredere ataşate parametrilor modelului se restrâng mult dacă valorile aberante nu sunt considerate în calcule.

Cel mai riguros mod de validare sau invalidare a unui model îl reprezintă aplicarea ***analizei de varianţă*** **(ANOVA)**. Aceasta reprezintă o tehnică statistică de analiză a măsurătorilor depinzând de mai mulţi factori care acţionează simultan în vederea estimării importanţei şi efectelor fiecăruia. Analiza de varianţă permite, cu alte cuvinte, să se determine ce parte a variabilităţii unei populaţii este cauzată de diverşi factori sistematici şi, respectiv, de întâmplare.

În continuare se prezintă metoda generală de validare a modelelor de regresie prin analiza de varianţă în cazul în care proprietatea de homoscedasticitate este îndeplinită şi presupunând că, în etapa de achiziţie a datelor, se repetă de n_i ori măsurarea răspunsului la fiecare combinaţie de valori a factorilor x_{ij}, pentru $j = 1, m$. Notăm cu $\hat{y}_i$ – media răspunsului la fiecare combinaţie de valori a factorilor şi cu $\bar{y}$ – media globală a valorilor determinate ale răspunsului în toate experimentele:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{n} \sum_{l=1}^{n_i} y_{il}$$

unde $N = \sum_{i=1}^{n} n_i$ este numărul total de experimente.

https://biblioteca-digitala.ro / https://unibuc.ro

Varianța valorilor y_{il} față de media globală, $\bar{y}$, este dată de suma de pătrate totală, SP_T:

$$SP_T = \sum_{i=1}^{n}\sum_{l=1}^{n_i}(y_{il} - \bar{y})^2$$

Această sumă poate fi descompusă în trei componente, deoarece se poate scrie ecuația următoare:

$$y_{il} - \bar{y} = (y_{il} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

, care introdusă în relația precedentă conduce, după efectuarea calculelor şi ținând cont de relațiile existente între variabile, la:

$$SP_T = \sum_{i=1}^{n}\sum_{l=1}^{n_i}(y_{il} - \bar{y}_i)^2 + \sum_{i=1}^{n} n_i(\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^{n} n_i(\hat{y}_i - \bar{y})^2$$

Tabelul 4.3
Descompunerea sumei de pătrate şi modul de validare al modelului de regresie prin analiza de varianță **(ANOVA)**

Sursa de variație	Suma de pătrate (SP)	Grade libertate	Media sumei de pătrate (MSP)*	Testul statistic F
Datorat regresiei	(SP_{REG}) $\sum_{i=1}^{n} n_i(\hat{y}_i - \bar{y})^2$	1	MSP_{REG}	
Variația mediei grupurilor față de modelul de regresie	(SP_{LC}) $\sum_{i=1}^{n} n_i(\bar{y}_i - \hat{y}_i)^2$	$n-2$	MSP_{LC}	$F = \frac{MSP_{LC}}{MSP_{EE}}$
Variația măsurătorilor repetate	(SP_{EE}) $\sum_{i=1}^{n}\sum_{l=1}^{n_i}(y_{il} - \bar{y}_i)^2$	$N-n$	MSP_{EE}	
Variația totală	(SP_T) $\sum_{i=1}^{n}\sum_{l=1}^{n_i}(y_{il} - \bar{y})^2$	$N-1$		

* Mediile sumelor de patrate se obtin facand catul dintre valorile din coloanele a doua si a treia.

https://biblioteca-digitala.ro / https://unibuc.ro

În tabelul 4.3 se prezintă semnificația termenilor sumei totale de pătrate împreună cu modul de efectuare al testului de validare a modelului matematic de regresie prin analiza de varianță.

Ultimul termen al sumei, SP_{REG}, este de obicei mare comparativ cu ceilalți doi termeni și exprimă variabilitatea valorilor estimate pentru variabila independentă, y, pe baza modelului de regresie față de media globală, $\bar{y}$.

Primul termen al sumei, SP_{EE}, exprimă variabilitatea măsurătorilor repetate și este afectat de $(N - n)$ grade de libertate. Media sumei de pătrate, MSP_{EE}, definită conform relației din tabel, reprezintă o estimație pură a erorilor experimentale, σ^2.

Termenul al doilea al sumei de pătrate, notat SP_{LC} și denumit sumă de pătrate datorată lipsei de concordanță, exprimă variabilitatea mediilor valorilor obținute prin repetarea experimentului față de modelul matematic de regresie și este afectat de $n - 2$ grade de libertate. Media acestui termen, MSP_{LC}, reprezintă și ea o estimație a varianței erorilor experimentale, σ^2, cu condiția ca modelul teoretic pentru efectuarea regresiei să fie adecvat. Așadar, validarea sau invalidarea modelului de regresie se poate face prin compararea statistică a egalității celor două medii a sumei de pătrate, MSP_{LC} și MSP_{EE}, utilizând testul statistic F, așa după cum rezultă din tabelul **ANOVA**. Dacă $F < F^{\alpha}_{n,N-n}$, unde n și $N - n$ reprezintă gradele de libertate corespunzătoare iar α este nivelul de semnificație al testului, atunci se poate admite, cu probabilitatea $P = 1 - \alpha$, că modelul de regresie utilizat exprimă corect relația deterministă dintre variabilele dependentă și independente.

4.2.4. Calculul intervalelor de încredere

Informația necesară determinării intervalelor de încredere a parametrilor este conținută în matricea de varianță – covarianță:

$$V(\boldsymbol{b}) = \begin{pmatrix} v(b_0) & \operatorname{cov}(b_0, b_1) & \operatorname{cov}(b_0, b_m) \\ \operatorname{cov}(b_1, b_0) & v(b_1) & \operatorname{cov}(b_1, b_m) \\ & & \cdot \\ \operatorname{cov}(b_m, b_0) & & v(b_m) \end{pmatrix}$$

https://biblioteca-digitala.ro / https://unibuc.ro

unde $v(b_i)$ reprezintă varianța parametrilor, b_i, considerând că aceştia reprezintă variabile aleatoare iar $cov(b_i, b_j)$ – covarianța parametrilor b_i şi b_j. Se poate arăta că aceasta are expresia:

$$V(\boldsymbol{b}) = s_{ee}^2 (X^T X)^{-1}$$

în care s_{ee}^2 reprezintă dispersia erorilor experimentale. În cazul în care aceasta nu este cunoscută iar modelul de regresie a fost validat, ea se poate înlocui cu expresia:

$$s_{ee}^2 = \frac{SP_e}{n-m-1}$$

în care SP_e reprezintă suma pătratelor tuturor reziduurilor iar n şi m au semnificațiile anterioare. În acest caz, expresia intervalului de încredere pentru oricare din parametri b_j, ($j = 0, m$), este:

$$b_j \pm t_{n-m-1}^{1-\alpha/2} \sqrt{v(b_j)}$$

unde $t_{n-m-1}^{1-\alpha/2}$ este valoarea testului Student pentru $n - m - 1$ grade de libertate şi pentru un nivel de semnificație α, iar $v(b_j)$ – valoarea elementului de pe linia $j + 1$ a matricii pătrate de varianță – covarianță, $V(\boldsymbol{b})$.

Trebuie subliniat faptul că deşi ultima ecuație este perfect valabilă pentru oricare parametru individual, ea nu redă corect expresia domeniului de încredere simultan pentru doi sau mai mulți parametri decât în cazul în care covariațele acestora, considerați rând pe rând doi câte doi, sunt egale cu zero.

4.3. Generalități privind determinările multicomponent

Există numeroase situații în care se pot determina simultan doi sau mai mulți compuşi dintr-o probă utilizând valorile de absorbanță din spectrele înregistrate în domeniile vizibil, ultraviolet şi/sau infraroşu. Aplicații de acest tip au fost propuse frecvent pentru controlul produselor farmaceutice în substanțele active utilizând metode de regresie multivariabile. Totuşi, caracteristicile de zgomot ale detectorilor spectrofotometrici şi corelația puternică dintre absorbanțele măsurate la lungimi de undă apropiate complică modul de selecție al variabilelor şi procesarea datelor.

Ca soluții la problemele menționate, literatura de specialitate descrie numeroase tipuri de regresii, acestea variind între modelele de regresie „***hard***" şi, respectiv,

https://biblioteca-digitala.ro / https://unibuc.ro

„soft". Modelele ***hard***, aşa cum este *regresia multivariabilă clasică*, **MLR**, sunt deduse direct din legile fizice pe care se bazează procesul de măsurare. Dimpotrivă, modelele ***soft*** (ca, de exemplu, *regresia prin metoda celor mai mici pătrate parţiale*, **PLS**, *reţelele neuronale artificiale*, **ANN**, *regresia ridge*, **RR** etc) se construiesc pornind de la datele brute măsurate experimental, cu puţină referinţă la cunoaşterea teoretică a sistemului chimic. De cele mai multe ori, relaţia deterministă dintre cauză şi efect se reflectă invers decât în realitate prin considerarea concentraţiilor de analiţi (care sunt controlate de către analist în soluţiile din setul de calibrare) ca răspuns al sistemului (sau variabile dependente) şi considerând caracteristicile măsurate (ca, de exemplu, absorbanţele la diferite lungimi de undă) ca variabile independente.

Procedând astfel în regresia **MLR**, concentraţiile analiţilor din probele de calibrare sunt ordonate într-o matrice $\boldsymbol{Y}$ cu n linii şi cu c coloane (n – numărul de probe din setul de calibrare; c – numărul de compuşi absorbanţi în domeniul spectral studiat, sau analiţi) iar valorile măsurate ale caracteristicilor pentru aceleaşi probe – într-o matrice $\boldsymbol{X}$ cu n linii şi p coloane (p – numărul de caracteristici spectrale măsurate, sau lungimi de undă, $p >> c$). Prin similitudine cu ecuaţia 4.3, se scrie relaţia:

$$\boldsymbol{Y}_{n,c} = \boldsymbol{X}_{n,p}\boldsymbol{B}_{p,c} + \boldsymbol{E}_{n,c} \qquad 4.9$$

, unde $\boldsymbol{B}_{p,c}$ reprezintă o matrice de parametri iar $\boldsymbol{E}_{n,c}$ – o matrice de erori. Prin algoritmul celor mai mici pătrate, matricea $\boldsymbol{B}_{p,c}$ se estimează prin:

$$\hat{\boldsymbol{B}}_{p,c} = (\boldsymbol{X}^T_{p,n}\boldsymbol{X}_{n,p})^{-1}\boldsymbol{X}^T_{p,n}\boldsymbol{Y}_{n,c} \qquad 4.10$$

, ecuaţie similară cu relaţia 4.6. Matricea $\hat{\boldsymbol{B}}_{p,c}$ se utilizează apoi, în etapa de predicţie, pentru a estima concentraţiile analiţilor dintr-o probă nouă, $\boldsymbol{y}_{0,1c}$, cu ecuaţia:

$$\hat{\boldsymbol{y}}_{0,1c} = \boldsymbol{x}_{0,1p}\hat{\boldsymbol{B}}_{p,c} \qquad 4.11$$

, unde $\boldsymbol{x}_{0,1p}$ este un vector linie ce conţine valorile caracteristicilor măsurate experimental pentru proba analizată.

Deşi, d.p.d.v. formal, valabilitatea relaţiei 4.9 este garantată de legile fizice liniare respectate în procesul de măsurare, ecuaţiile 4.10 şi 4.11 trebuie privite cu rezerve datorită faptului că ipotezele statistice de independenţă a variabilelor presupuse ca independente şi de distribuţie homoscedastică, pe care se bazează deducerea ecua-

https://biblioteca-digitala.ro / https://unibuc.ro

ției 4.10, nu pot fi simplu testate și, de cele mai multe ori, nici nu sunt îndeplinite. De aceea, problemele specifice regresiilor (de identificare a probelor anormale - *outliers*, de selecție a variabilelor independente cu cea mai mare semnificație și, respectiv, de evaluare a erorilor conținute în estimațiile obținute) se complică foarte mult. Din acest motiv, analogia constatată anterior cu regresia liniară monovariabilă nu se conduce, de cele mai multe ori, până la ultimele sale limite. În loc de aceasta, elaborarea unui model de calibrare acceptabil se face utilizând subseturi multiple de probe de calibrare și un procedeu de validare specific, denumit ***validare încrucișată***. Dacă probele de analizat sunt asemănătoare cu probele utilizate la calibrare, se acceptă ipoteza că mediile erorilor obținute pentru concentrațiile de analiți cu diferite modele de regresie, aplicate însă asupra unor subseturi de probe de calibrare diferite de cele utilizate la elaborarea modelelor, reprezintă o estimație bună a erorilor de predicție a concentrațiilor de analiți din probele analizate.

Între terenul stâncos al modelelor ***hard*** și nisipurile mișcătoare ale modelelor ***soft***, un loc de mijloc este ocupat de către *regresia după componete principale* a lui Massy, **PCR**. În acest caz, caracteristicile măsurate inițial pentru probele utilizate la calibrare se utilizează mai întâi pentru a calcula un sistem de coordonate propriu sistemului chimic studiat și procesului de măsurare utilizat. *Componentele principale* reprezintă acum axele noului sistem de coordonate rezultat și, prin construcție, acestea sunt perpendiculare între ele două câte două și ordonate în funcție de valorile descrescătoare ale variațelor de-a lungul lor al caracteristicilor măsurate experimental. De obicei, cea mai mare parte a variației totale a caracteristicilor măsurate și datorată variabilității concentrațiilor de analiți din probele de calibrare se repartizează de-a lungul unui număr redus de componente principale, considerate semnificative, în timp ce restul componentelor principale cumulează în principal variață datorată zgomotului din sistem și, prin urmare, pot fi ignorate. În felul acesta, rezultă în final un model MLR mult mai simplu folosind ca noi variabile independente „proiecțiile" valorilor determinate ale caracteristicilor pe componentele principale semnificative. Procedând astfel, **PCR** utilizează în etapa de predicție toată informația conținută în caracteristicile determinate pentru probele necunoscute sub forma unui număr redus de variabile

https://biblioteca-digitala.ro / https://unibuc.ro

necorelate între ele. Ca rezultat, varianța de predicție este mai mică decât în cazul regresiei **MLR** clasice.

Dezavantajul principal al **PCR** constă în faptul că noile variabile obținute (componentele principale) se calculează fără nici o referire la concentrațiile de analiți din probele de calibrare. În consecință, este posibil ca *„răspunsul sistemului"* să nu fie suficient de bine descris de modelul PCR (*underfit*), în special în cazul unor calibrări proiectate defectuos. De aceea, cei mai mulți chimiști analiști utilizează regresia PLS a lui Wold pentru a construi, prin algoritmi de calcul foarte complicați, așa numitele ***variabile latente*** (corespondentele componentelor principale din **PCR**), mai bine corelate cu răspunsul sistemului. Oricum, s-a arătat (Höskuldsson - 1992) că **MLR** și **PCR** reprezintă două cazuri extreme ale **PLS**, aceasta din urmă reprezentând optimul din punctul de vedere al **problemei** *H* a regresiilor multivariabile. Aceasta (problema *H*) este similară principiului lui Heisenberg din fizică și reflectă faptul că, atunci când elaborarea modelului de regresie depinde foarte mult de (sau, metaforic, interacționează puternic cu) structura variabilelor independente, o precizie mare la estimarea parametrilor în etapa de calibrare se opune unei varianțe minime a estimațiilor obținute în etapa de predicție.

4.4. Concluzii

În acest capitol, s-au prezentat algoritmii generali de efectuare a celor mai uzuale tipuri de regresii folosite în analiza chimică și s-au sugerat câteva domenii de aplicare ai acestora. Problemele generale ale regresiilor multicomponent au fost doar enunțate succint deoarece există în literatura recentă numeroase lucrări care ilustrează aprofundat suportul teoretic al acestora.

Deși algoritmii de regrsie prezentați pot părea complicați la prima vedere, aceștia de derulează simplu cu ajutorul calculatoarelor personale, existând un mare număr de programe specializate pentru acest scop.

Bibliogarafie

1. C. Liteanu și I. Rîcă, ***Teoria și metodologia statistică a analizei urmelor***, Editura Scrisul Românesc, Craiova, 1979.

https://biblioteca-digitala.ro / https://unibuc.ro

2. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte şi L. Kaufman, ***Chemometrics: A. Textbook***, Elsevier, Amsterdam, 1988.

3. R. Shiavi, ***Introduction to Applied Statistical Analysis***, Aksen Publishers, Boston, 1991.

4. S. D. Brown şi R. S. Bear, Jr., *Crit. Rev. Anal. Chem.*, **24**, 99, 1993.

5. E. V. Thomas, *Anal. Chem.*, **66**, 795A, 1994.

6. I. Gh. Şabac; P. Cocârlan; O. Stănăşilă şi A. Topală, ***Matematici Speciale***, vol. II, E. D. P., Bucureşti, 1983.

7. C. Liteanu şi I. Rîcă, ***Optimizarea proceselor analitice***, Editura Academiei Române, Bucureşti, 1985.

8. N. R. Draper şi H. Smith, ***Applied Regression Analysis***, Wiley, New York, 1981.

9. D. Ceauşescu, ***Utilizarea statisticii matematice în chimia analitică***, Editura Tehnică, Bucureşti, 1982.

10. K. Ebert; H. Ederer şi T. Lisenhour, ***Computer Applications in Chemistry. An Introduction for PC Users***, VCH, 1989.

11. S. Wolfram, ***Mathematica™ A. System for Doing Mathematics by Computer***, Addison-Wesley, New York, 1988.

12. R. Jain şi J. S. Schultz, *Anal. Chem.*, **56**, 141, 1984.

13. J. F. Rusling, *Crit. Rev. Anal. Chem.*, **21**, 49, 1989.

14. J. H. Kalivas, *Anal. Chim. Acta*, **428**, 31, 2001.

15. M. C. Ortiz Fernandez, *"Modelos de regression lineal para datos experimentales"*, în *** ***Curso COMETT II - EUROCHEMOMETRICS***, Burgos, 1993.

16. V. David, I. G. David şi G. L. Radu, *Rev. Chim. (Bucharest)*, **54**, 97, 2003.

17. W. F. Massy, *J. Amer. Stat. Assoc.*, **60**, 234, 1965.

18. H. Wold, *„Soft Modeling by Latent Variables...”*; în ***Perspectives in Probability and Statistics***, J. Gani (editor), Academic Press, Londra, 1975.

19. M. Otto, ***Chemometrics, Statistics and Computer Application in Analytical Chemistry***, Wiley – VCH, Weinheim, 1998.

20. A. Höskuldsson, *Chemom. Intel. Lab. Syst.*, **14**, 139, 1992.

https://biblioteca-digitala.ro / https://unibuc.ro

5 Optimizarea şi proiectarea experimentelor

5.1 Aspecte ale optimizării în chimia analitică

O preocupare frecventă a analiştilor este aceea de a adapta o metodă existentă unui anumit tip de probă sau unui anumit laborator. De exemplu trebuie ajustat volumul de reactiv, timpul de reacţie sau temperatura în vederea atingerii performanţelor corespunzătoare problemei studiate deci trebuie optimizată metoda.

Optimizarea nu presupune numai modificarea parametrilor pentru obţinerea optimului metodei ci şi modificarea însăşi a sistemului. Problema alegerii optime a unei metode sau a unui grup de metode pe baza unor informaţii existente în literatură poate fi rezolvată cu ajutorul chemometriei. Deci ca o primă etapă în cadrul procesului de optimizare a unei metode de analiză o constituie alegerea caracteristicii sau caracteristicilor de performanţă cu ajutorul cărora urmează ulterior să se definitiveze deciziile corespunzătoare.

In cadrul metodelor utilizate în analiza chimică se pot defini o serie de caracteristici după cum urmează: caracteristici tehnice (precizie, corectitudine, selectivitate, limită de detecţie, cantitatea de informaţie, amplitudinea semnalului, raportul semnal/zgomot, eroarea totală de analiză, limita de determinare, rezoluţia, etc.) şi caracteristici economice (costul, timpul necesar efectuării unei determinări, etc.

5.1.1 Reprezentarea matematică a sistemului de optimizat

Modelul matematic al unui proces este dat de relaţia dintre variabila de răspuns (**y**) şi variabilele independente ($x_1, x_2, \ldots\ldots x_n$) model prezentat în ecuaţia de mai jos:

$$y = f(x_1, x_2, \ldots\ldots x_n, b_1, b_2, \ldots\ldots b_n) \qquad (5.1)$$

în care $b_1, b_2, \ldots\ldots\ldots b_n$ reprezintă parametrii modelului.

In cazul în care procesul nu poate fi modelat matematic adecvat se folosesc metodele statistice ale experimentului active şi anume experimentul activ se desfăşoară conform unui plan experimental elaborat în prealabil.

https://biblioteca-digitala.ro / https://unibuc.ro

Reprezentarea matematică adecvata a sistemului de optimizat constituie obiectivul principal în rezolvarea optimizării privind dependența dintre criteriul de optimizare şi variabilele de optimizat. Problema optimizării unui sistem, în general este abordată după descompunerea sistemului în subsistemele componente deoarece este mai uşor de rezolvat optimizarea pe subsisteme datorită numărului mai redus de variabile. In general, un sistem analitic poate fi descompus în următoarele subsisteme: eşantionarea materialului, tratamentul fizico-chimic al eşantionului, măsurarea semnalului analitic, prelucrarea semnalelor şi calculul rezultatelor. Sistemul analitic cu funcție de detecție, identificare sau determinare este un lanț informațional în care la intrare se găseşte materialul de analizat iar la ieşire rezultatul de analiză.

In cazul reprezentării matematice a unui sistem analizat în scopul optimizării acestuia se desting două cazuri [1]:

- sistemul de optimizat este riguros cunoscut deci poate fi reprezentat cantitativ în vederea evaluării funcției obiectiv (dependența dintre criteriul de optimizat şi variabilele sistemului) cât şi a restricțiilor pe care trebuie să le respecte funcția obiectiv şi variabilele sistemului;

- sistemul de optimizat există în funcțiune iar reprezentarea sa nu este posibilă sau este greu de realizat. In acest caz, funcția obiectiv se evaluează pe cale experimentală în acord cu principiul black-box.

5.1.2 Funcția de răspuns şi restricțiile

Funcția care exprimă dependența dintre caracteristica de calitate (criteriu de optimizare) şi variabilele de optimizat se numeşte funcție de răspuns (funcție obiectiv sau funcție de optimizat).

In general, funcția de răspuns se notează prin expresia de mai jos:

$$y(X) = y(x_1, x_2, \ldots\ldots\ldots x_n) \qquad (5.2)$$

https://biblioteca-digitala.ro / https://unibuc.ro

In practică, variabilelor $x_1, x_2, \ldots\ldots x_n$ le sunt impuse de regulă restricții, adică ele nu pot lua valori decât în anumite domenii. Se disting restricții de tip egalitate notate prin:

$$g_i = g(X) = 0, \quad i = 1, 2, \ldots\ldots m \tag{5.3}$$

și restricții de tip inegalitate prin:

$$h_j = x(X) > 0, \quad j = 1, 2, \ldots\ldots l \tag{5.4}$$

In chimia analitică funcțiile obiectiv pot fi: eroarea de analiză, timpul afectat unei determinări, capacitatea de detecție, costul unei determinări, determinarea simultană a mai multor componenți dintr-o probă.

Funcția obiectiv în cazul optimizării capacității de detecție și determinare a unei metode este de fapt raportul semnal/zgomot. Deci trebuiesc alese acele condiții analitice cărora le corespunde un raport semnal/zgomot maxim. In general, problema optimizării răspunsului în chimia analitica se rezumă la găsirea acelor valori pentru condițiile analitice pentru care răspunsul capătă o valoare maximă.

Alegerea metodei de optimizare este determinată și de complexitatea matematică a funcției obiectiv și a restricțiilor. Chiar pentru sisteme relativ simple, funcția de răspuns (obiectiv) și restricțiile pot avea formulări matematice deosebit de complicate. Pentru rezolvarea problemelor de optimizare în acest caz sunt necesare metode numerice adecvate.

In cazul unei variabile de optimizat funcția obiectiv poate fi reprezentată grafic și examinată în vederea obținerii optimului. In acest caz se constată că optimul ales poate însemna o valoare bine determinată pentru variabila de optimizat sau un interval.

In cazul a două variabile x_1 și x_2 , funcția de răspuns este reprezentată în spațiul tridimensional de către o suprafață de forma următoare (figura 5.1):

$$y = y(x_1, x_2) \tag{5.5}$$

sau printr-un sistem de curbe de nivel constant $y = y(x_1, x_2) = \textbf{constant}$ (figura 5.2). In cazul a n variabile independente de optimizat ($n > 2$) funcția de răspuns este reprezentată de o hipersuprafață în spațiul $n+1$ dimensiuni. Hipersuprafețele $y = y(X) =$ **constant** sunt denumite hipersuprafețe de nivel constant al funcției de răspuns.

https://biblioteca-digitala.ro / https://unibuc.ro

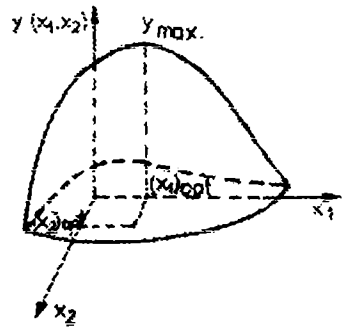


Figura 5.1. *Reprezentarea grafică a funcției obiectiv de două variabile.*

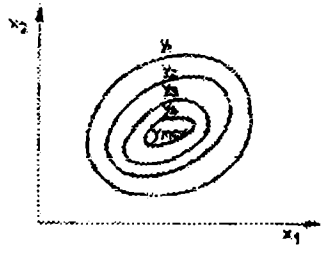


Figura 5.2 *Reprezentarea grafică a funcției obiectiv de două variabile prin curbe de nivel constant.*

Deci optimizarea se defineşte ca o cercetare a valorilor lui X_i care dau valoarea lui **Y** cea mai ridicată sau cea mai scăzută. Atât problema reprezentării matematice a sistemului de optimizat deci a stabilirii funcției obiectiv cât şi problema găsirii maximumului sau minimumului funcției nu sunt probleme simple de abordat.

$$\mathbf{opt.[y\ (X),\ldots g_i\ (X) = 0,\ i = 1,2,\ldots\ldots m;\ h_j\ (X) > 0,\ j = 1,2,\ldots l]} \qquad (5.6)$$

reprezintă maximul sau minimul funcției obiectiv **y(X)** şi restricțiile corespunzătoare. Trecerea de la o problemă de maxim la una de minim se face conform relației:

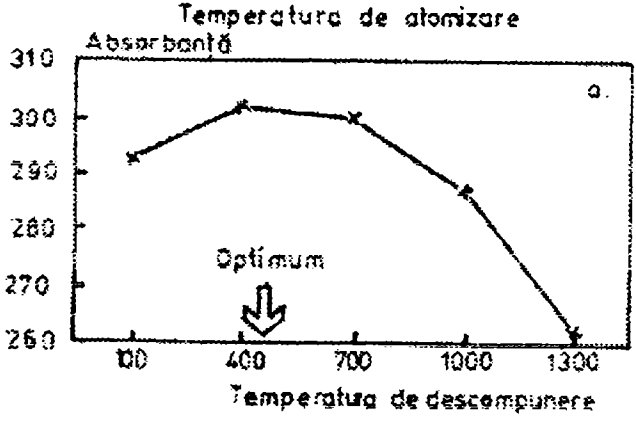
$$\mathbf{max.\ y\ (X) = min.\ [-y\ (X)]} \qquad (5.7)$$

In concluzie problema optimizării este deplasată către aceea a explorării suprafeței de răspuns.

https://biblioteca-digitala.ro / https://unibuc.ro

5.1.3 Strategia optimizării

In general punerea la punct a unei metode constă în modificarea unui singur factor, restul factorilor fiind menținuți constanți. In figura 5.3, ***a*** şi ***b*** este prezentată grafic această strategie aplicată în cazul spectrometriei de absorbție atomică fără flacără în tub de grafit. Obiectivul acestui experiment [2] este de a găsi temperaturile de descompunere şi respectiv de atomizare care dau o absorbanță maximă. Dacă se iau în considerare numai cei doi factori: temperatura de descompunere (TD) si temperatura de atomizare (TA) metodologia clasică constă de exemplu în aceea de a menține mai întâi factorul TA invariabil la o valoare convenabil aleasă şi a modifica factorul TD (figura 5.3, ***a***).Prin evaluarea răspunsurilor corespunzătoare se poate alege valoarea optimă pentru factorul TD, căreia îi corespunde o valoare maximă pentru răspuns. In continuare pentru a stabili valoarea optimă pentru factorul TA, se menține factorul TD invariabil la optimul găsit şi se modifică factorul TA (figura 5.3, ***b***).



https://biblioteca-digitala.ro / https://unibuc.ro

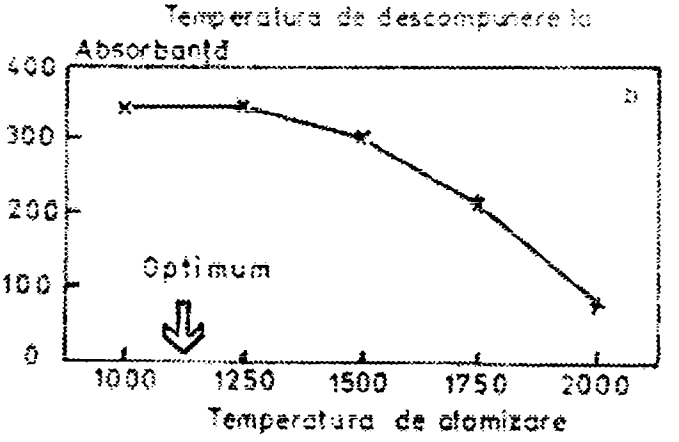


Figura 5.3, a, b *Studiul izolat al factorilor*

Dacă se reprezintă cele două serii de măsurători simultan (figura 5.4), simulând absorbanțe teoretice, se observă că optimul găsit aplicând metodologia variației unui singur factor , întâi TD apoi TA, nu corespunde optimului adevărat. Această eroare apare datorită faptului că metodologia clasică de a modifica un singur factor nu ține seama de interacțiunea care există între factorii TA şi TD.

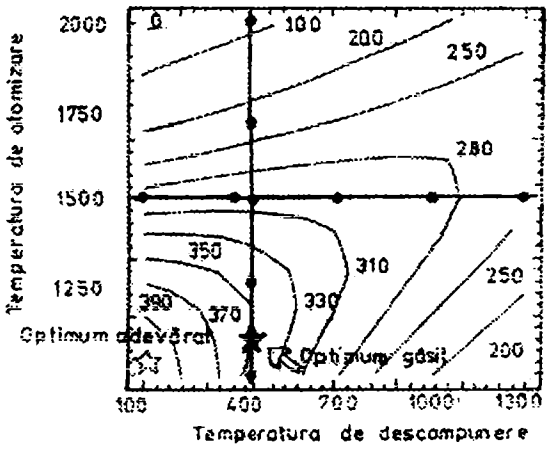


Figura 5.4 *Reprezentarea simultană a două serii de măsurători.*

https://biblioteca-digitala.ro / https://unibuc.ro

Evident au fost propuse şi alte strategii care ţin seama de interacţiile între factori. Una dintre ele ar fi cea deterministă în care valorile optime pentru variabila de optimizat se găsesc cu ajutorul metodelor de programare matematică. Aceste ecuaţii stabilite în baza legilor fizicii şi chimiei sunt rezolvate algebric în aşa fel încât sunt puse în evidenţă soluţiile cele mai favorabile. Altă strategie ar fi cea empirică şi experimentală care constă în modelarea metodei studiate ca o cutie neagră (black box) şi evaluarea comportamentului său cu ajutorul încercărilor planificate (figura 5.5). Este cazul metodelor planurilor experimentale [3] şi al metodei simplex [4]. Pentru cea mai mare parte a chimiştilor soluţia deterministă apare ca fiind cea mai comodă şi ea este adesea utilizată în procesul de optimizare. Insă, în cele mai multe cazuri ea nu poate fi aplicată deoarece nu întotdeauna fenomenul studiat poate fi evaluat teoretic riguros. De aceea este necesară de cele mai multe ori în special în cazul chimiei analitice o planificare a încercărilor pentru a evita erorile de interpretare.

Metoda optimizării este văzută de un experimentator ca un sistem închis din care se va selecţiona o ieşire notată cu **y**, numită criteriu de optimizare.

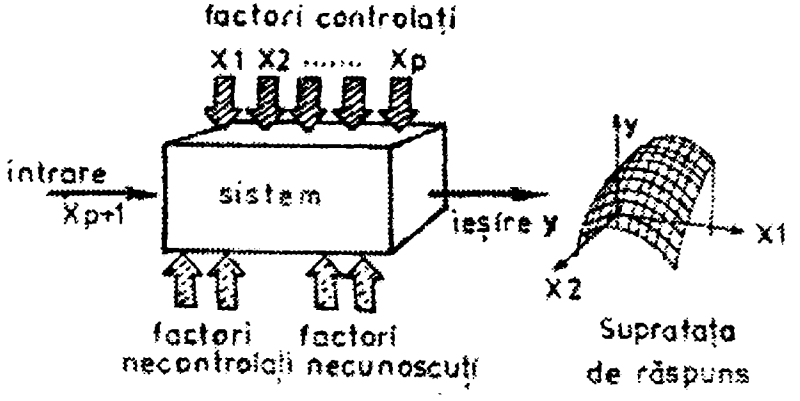


Figura 5.5 *Sistemul cutiei negre*

Se presupune că valorile **y**, variază datorită unui număr de factori controlaţi de experimentator care iau anumite valori fixe. Aceste valori fixe se numesc nivelele factorilor. Reprezentarea variaţiilor lui **y** în funcţie de nivelele factorilor formează

https://biblioteca-digitala.ro / https://unibuc.ro

suprafața de răspuns specifică sistemului studiat. Prin convenție se notează de obicei cu **p** factorii $x_1 \ldots\ldots x_p$ și nivelele lor pentru **n** încercări x_{11}, x_{21},...$x_{n-1,p}$, x_{np}. De exemplu, în cazul unei metode de dozare colorimetrică, ieșirea **y** va fi intensitatea reacției, măsura înălțimii picurilor, iar factorii controlați vor fi temperatura X_1, durata reacției X_2 și pH-ul, X_3. De asemenea, întotdeauna trebuie avut în considerație faptul că există factori cunoscuți necontrolați ca de exemplu: fluctuațiile lămpii colorimetrului sau evaporarea soluției în cuva de măsură precum și factori necunoscuți și necontrolați. Suma acestor factori necontrolați formează zgomotul de fond sau eroarea experimentală a metodei și se notează cu **e**.

Plecând de la aceste două concepte al factorilor controlați și al zgomotului de fond aleatoriu se poate scrie un model matematic astfel:

$$y = f(x_1, x_2, \ldots\ldots x_p) + e \qquad (5.8)$$

5.1.4 Metode de evaluare a optimului funcției obiectiv

Există mai multe tipuri de metode care pot fi aplicate la stabilirea optimului unei funcții obiectiv.

5.1.4.1 Metodele analitice sunt metodele care dau informații de ansamblu asupra funcției și ele presupun continuitatea funcției și a derivatelor ei. Aceste metode folosesc calculul diferențial pentru maximul sau minimul funcției. Totuși multe dintre sistemele de interes practic nu pot fi optimizate prin aceste metode.

5.1.4.2 Metodele numerice sunt metodele care utilizează un algoritm pentru stabilirea optimului funcției obiectiv. In acest caz funcția obiectiv este tratată ca un sistem black-box [1]. In cazul aplicării metodelor numerice la optimizarea unui proces analitic, se disting următoarele două situații:

- funcția de optimizat este cunoscută în formă matematică exactă, caz în care pentru evaluarea optimului , sistemul se tratează ca un sistem black-box;

- funcția de optimizat nu este cunoscută deci în acest caz se acționează asupra sistemului de optimizat aflat în funcțiune, acesta fiind tratat ca un sistem black-box.

https://biblioteca-digitala.ro / https://unibuc.ro

5.1.4.3 Metodele mixte sunt metodele în cazul cărora funcția obiectiv se tratează ca un sistem black-box. In baza unui plan experimental se face evaluarea numerică a funcției obiectiv pentru mai multe valori impuse variabilelor de optimizat. Cu ajutorul acestor date se poate rezolva analitic funcția obiectiv în scopul stabilirii optimului.

Metodele prezentate mai sus pot fi aplicate unui sistem riguros cunoscut. Dar cazurile în care se cunoaşte sau se poate deduce teoretic funcția de răspuns cât şi modelul matematic pentru aceasta sunt limitate. In majoritatea cazurilor în chimia analitică atât valoarea funcției de răspuns cât şi modelul matematic pentru aceasta nu sunt cunoscute. In acest caz pentru rezolvarea problemelor de optimizare se aplica metodele experimental-statistice [5] conform cărora sistemul în funcțiune este studiat în acord cu principiul black-box.

Sistemul de optimizat există în funcțiune astfel încât se modifică controlat variabilele de optimizat evaluându-se răspunsul sistemului pentru fiecare stare nouă impusă variabilelor. Metodele de optimizare experimental-statistice sunt descrise în numeroase lucrări de specialitate care tratează probleme teoretice şi practice în această direcție[6-9].

Există două posibilități de abordare a metodelor de optimizare experimental-statistice şi anume:

- una dintre posibilități consideră volumul de valori experimentale succesive (acestea cuprind valori ale variabilelor de optimizat cât şi valori ale răspunsului sistemului, corespunzătoare) şi dirijează cercetarea în direcția optimului;

- a doua posibilitate constă în desfăşurarea optimizării după un plan experimental stabilit prealabil. Evident planul experimentului se poate modifica pe parcursul desfăşurării experimentului.

Planul experimental este conceput astfel încât cu ajutorul mulțimii de date experimentale să se poată evalua răspunsul sistemului cât mai adecvat. Valorile optime pentru variabilele sistemului sunt prelucrate prin cercetare numerică sau analitică. Pentru optimizarea răspunsului în chimia analitică adesea este utilizată metoda simplex,

https://biblioteca-digitala.ro / https://unibuc.ro

metodă elaborată inițial de Spendley şi colab.[10] şi apoi modificată de Nedler şi Mead [11].

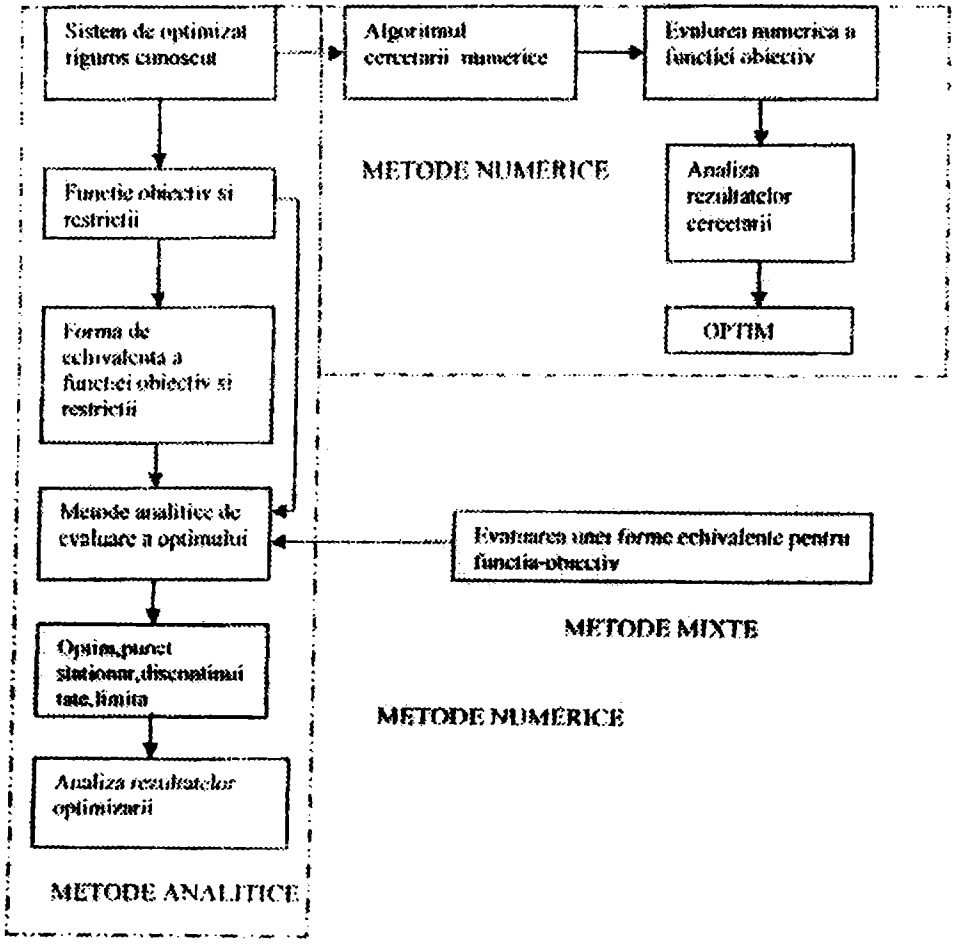
Metoda simplex constă într-un algoritm iterativ de localizare a optimului funcției obiectiv pe suprafața de răspuns.

Cele două strategii de optimizare şi anume cea a planului experimental şi simplex au o aplicație largă în chimia analitică. Utilizarea unui plan experimental pentru explorarea unei suprafețe de răspuns şi găsirea optimului se face în 6 etape a căror secvență trebuie respectată:

1. Alegerea unui criteriu de optimizare
2. Selecția factorilor controlați şi delimitarea exactă a domeniului experimental
3. Alegerea unui model de preferință liniar
4. Alegerea unui plan experimental
5. Obținerea determinărilor, calculul coeficienților şi localizarea optimului practic
6. Efectuarea unei determinări complementare a optimului pentru validarea modelului.

Mai jos este prezentată o schemă generală a metodelor utilizate pentru stabilirea optimului unei funcții obiectiv.

https://biblioteca-digitala.ro / https://unibuc.ro



5.2 Metodologia suprafeței răspunsului

5.2.1 Suprafața de răspuns

Suprafața de răspuns este dată de reprezentarea grafică a răspunsului sistemului în funcție de unul sau mai mulți factori sau variabile independente. Suprafața de răspuns oferă chemometristului posibilitatea de a vizualiza modul în care este afectată măsurarea sistemului studiat de către diferiți factori. Studiul metodologiei suprafetei răspunsului a fost descrisă şi dezvoltată pentru prima dată de către Box şi Wilson[12, 13].

https://biblioteca-digitala.ro / https://unibuc.ro

Metodologia suprafeței răspunsului este o strategie experimental-statistică aplicată studiului răspunsului sistemelor. În general această metodologie se aplică în studiul proceselor industriale evident pentru îmbunătățirea performanțelor acestora dar sunt și posibilități de utilizare în chimia analitică privind răspunsul proceselor chimice.

Considerăm funcția de răspuns necunoscută în forma generală:

$$\hat{y} = f(x_1, x_2, \ldots x_n) \tag{5.9}$$

Dacă se dezvoltă funcția de răspuns în serie Taylor, în jurul unui punct, rezultă că orice funcție se poate aproxima printr-un polinom de ordinul I:

$$y = b_0 + \sum_{i=1}^{n} b_i x_i \tag{5.10}$$

de ordinul II:

$$y = b_0 + \sum_{i=1}^{n} b_i x_i + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} b_{ij} x_i x_j + \sum_{i=1}^{n} b_{ii} x_i^2 \tag{5.11}$$

sau printr-un polinom de ordin superior. În concluzie, oricât de complicată ar fi funcția de răspuns, aceasta poate fi aproximată printr-un model liniar în raport cu parametrii pe care îi conține.

În acest caz, estimarea coeficienților prin metoda celor mai mici pătrate se face din condiția de minim pentru suma pătratelor deviațiilor (SPD):

$$SPD = \sum_{i=1}^{n} \left(y_i - b_0 + b_1 x_{1i} + \ldots + b_{11} x_{1i}^2 + \ldots \right)^2 = \text{minim} \tag{5.12}$$

adică, din sistemul de ecuații normale:

https://biblioteca-digitala.ro / https://unibuc.ro

$$\frac{\partial (SPD)}{\partial b_0} = 0$$
$$\frac{\partial (SPD)}{\partial b_1} = 0 \qquad (5.13)$$
$$\cdots\cdots\cdots\cdots$$
$$\frac{\partial (SPD)}{\partial b_{11}} = 0$$

În cazul polinomului de ordinul întâi având următoarea expresie:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \qquad (5.14)$$

sistemul de ecuații normale obținut cu ajutorul metodei celor mai mici pătrate în care coeficienții modelului $\beta_0, \beta_1, \ldots, \beta_k$ au fost înlocuiți cu parametrii de estimare $b_0, b_1, \ldots, b_k$ are expresia de mai jos:

$$b_0 \sum_{i=1}^{n} x_{0i}^2 + b_1 \sum_{i=1}^{n} x_{0i} x_{1i} + \ldots + b_k \sum_{i=1}^{n} x_{0i} x_{ki} = \sum_{i=1}^{n} x_{0i} y_i$$
$$b_0 \sum_{i=1}^{n} x_{1i} x_{0i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + \ldots + b_k \sum_{i=1}^{n} x_{1i} x_{ki} = \sum_{i=1}^{n} x_{1i} y_i$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$b_0 \sum_{i=1}^{n} x_{ki} x_{0i} + b_1 \sum_{i=1}^{n} x_{ki} x_{1i} + \ldots + b_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} x_{ki} y_i$$

Acest sistem de ecuații poate fi reprezentat compact sub formă matricială:

$$\mathbf{X^T X \beta = X^T y} \qquad (5.15)$$

https://biblioteca-digitala.ro / https://unibuc.ro

unde am introdus matricea X şi transpusa sa $\mathbf{X^T}$ pentru n experimente şi k variabile independente:

$$\mathbf{X} = \begin{Vmatrix} x_{01} & x_{11} & \cdots & x_{k1} \\ x_{02} & x_{12} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{0n} & x_{1n} & \cdots & x_{kn} \end{Vmatrix}, \tag{5.16}$$

matricea valorilor răspunsurilor corespunzătoare fiecărui experiment reprezentat printr-o linie a matricii X:

$$\mathbf{y} = \begin{Vmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{Vmatrix} \tag{5.17}$$

precum şi matricea coeficienţilor polinomului de aproximare:

$$\boldsymbol{\beta} = \begin{Vmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_k \end{Vmatrix} \tag{5.18}$$

Atunci, în condiţiile în care matricea $\mathbf{X^T X}$ este inversabilă (determinantul său asociat de ordinul k+1 nenul), rezolvând ecuaţia (5.17) în raport cu matricea β se obţin valorile de estimare:

$$\beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{5.19}$$

https://biblioteca-digitala.ro / https://unibuc.ro

5.2.2 Experimente factoriale

Dacă în cazul unui experiment activ, matricea X de programare a experimentului se poate defini astfel încât matricea $\mathbf{X^T X}$ să devină o matrice diagonală de forma prezentată mai jos:

$$\mathbf{X^T X} = \begin{Vmatrix} \sum_{j=1}^{n} x_{0j}^2 & 0 & \cdots\cdots & 0 \\ 0 & \sum_{j=1}^{n} x_{1j}^2 & \cdots\cdots & 0 \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ 0 & 0 & \cdots\cdots & \sum_{j=1}^{n} x_{kj}^2 \end{Vmatrix} \qquad (5.20)$$

acest experiment cu o matrice $\mathbf{X^T X}$ de formă diagonală se numeşte experiment ortogonal [14, 5]. In acest caz, coeficienţii polinomului (5.10) se calculează astfel:

$$b_i = \left(\sum_{j-1}^{n} x_{ij}^2 \right)^{-1} \sum_{j-1}^{n} x_{ij} y_i \qquad (5.21)$$

unde i=0,1,2,.....k şi j=1,2,.....n.

Ca un avantaj esenţial al utilizării experimentului ortogonal la studiul suprafeţei de răspuns o constituie faptul că asigură un volum de calcule mic pentru evaluarea coeficienţilor polinomului şi analiza statistică a rezultatelor experimentale. Coeficienţii obţinuţi cu ajutorul unui experiment factorial ortogonal nu sunt interdependenţi şi deci în cazul corecturii modelului matematic, coeficienţii deja calculaţi rămân valabili.

https://biblioteca-digitala.ro / https://unibuc.ro

5.2.2.1 Experiment factorial complet la două nivele

(Plan experimental 2^n)

Experimentele factoriale sunt o clasă de experimente foarte economice care oferă o cantitate mare de informație dintr-un număr mic de experimente. Când numărul experimentelor efectuate sunt în număr mic, experimentul factorial oferă calea cea mai eficientă pentru a obține maximul de informație al experimentelor. Experimentele factoriale pot fi de două feluri:

1) de tipul 2^n unde 2 reprezintă numărul de nivele iar n, numărul factorilor;

2) de mai multe nivele, cu 2 sau mai mulți factori.

Cea mai cunoscută schemă de programare a experimentului factorial este aceea în care variabilele (factorii) iau numai două valori de o parte și de alta a unui nivel de bază. În chimia analitică factorii pot fi: pH, temperatura, timpul, reactivi utilizați, aparat de măsură, mod de lucru, etc. Se poate programa experimental și la mai multe nivele dar în acest caz realizarea experimentului și prelucrarea datelor sunt mult mai grele. În cazul unui experiment factorial la două nivele, pentru toți factorii, numărul de combinații posibile este egal cu 2^k. În acest caz avem un experiment factorial cu un plan experimental 2^k. Se notează nivelul inferior și superior pentru fiecare factor în parte prin $(x_i)_1$ și $(x_i)_2$, unde i = 1,2,.....n și se introduc pentru fiecare variabilă în parte mărimile:

$$x_i^0 = \frac{(x_i)_1 + (x_i)_2}{2}, \quad \text{nivel de baza}$$

$$\Delta x_i = \frac{(x_i)_2 - (x_i)_1}{2}, \quad \text{unitatea intervalului} \qquad (5.22)$$

$$x_i' = \frac{x_i - x_i^0}{\Delta x_i}, \quad \text{valori codificate}$$

https://biblioteca-digitala.ro / https://unibuc.ro

Deci, în urma centrării şi normării, pentru fiecare factor în parte, nivelul superior este codificat cu **+1** iar nivelul inferior cu **-1**, coordonatele centrului fiind nule. În cazul particular a doi factori, k=2 deci $\mathbf{n = 2^2 = 4}$; toate combinaţiile posibile între nivelele celor doi factori sunt date în figura 5.6.

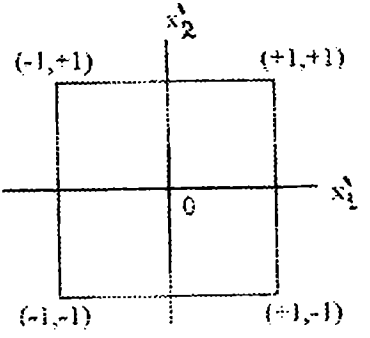


Figura 5.6 *Experiment factorial complet pentru doi factori la două nivele*

Matricea de programare a unui experiment complet pentru doi factori la două nivele fiecare este prezentată în tabelul 5.1.

Tabel 5.1 *Programul experimentului factorial* 2^2

Număr experiment	Factori			Răspuns
	$\mathbf{x'_0}$	$\mathbf{x'_1}$	$\mathbf{x'_2}$	
1	+1	+1	+1	$\mathbf{y_1}$
2	+1	-1	+1	$\mathbf{y_2}$
3	+1	+1	-1	$\mathbf{y_3}$
4	+1	-1	-1	$\mathbf{y_4}$

https://biblioteca-digitala.ro / https://unibuc.ro

În cazul a trei factori, numărul de combinații posibile între toți factorii pentru două nivele $n = 2^3 = 8$ după cum se poate vedea în figura 5.7 iar matricea de programare a unui experiment complet la două nivele pentru trei factori este prezentată în tabelul 5.2.

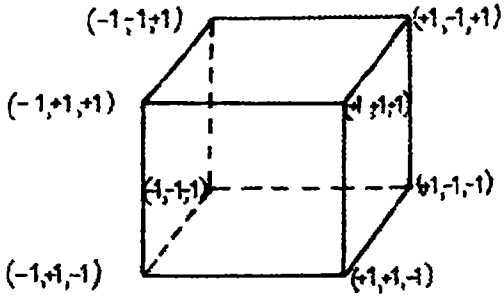


Figura 5.7 *Experiment factorial complet pentru 3 factori la două nivele.*

Tabelul 5.2 *Experiment factorial* 2^3

Număr experiment	Factori				Răspuns
	x'_0	x'_1	x'_2	x'_3	
1	+1	+1	+1	+1	y_1
2	+1	-1	+1	+1	y_2
3	+1	+1	-1	+1	y_3
4	+1	-1	-1	+1	y_4
5	+1	+1	+1	-1	y_5
6	+1	-1	+1	-1	y_6
7	+1	+1	-1	-1	y_7
8	+1	-1	-1	-1	y_8

https://biblioteca-digitala.ro / https://unibuc.ro

În acest caz, valorile de estimare $\mathbf{b_0, b_1, \ldots, b_n}$, pentru coeficienții polinomului liniar conform relației 5.21 se calculează astfel:

$$B = \begin{Vmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{Vmatrix} = \left(X^T X\right)^{-1} X^T Y \qquad (5.23)$$

adică:

$$\mathbf{b_i = \frac{1}{n} \sum_{j=1}^{n} x'_{ij} y_j, \quad i = 0,1,\ldots,k; \quad j = 1,2,\ldots,n} \qquad (5.24)$$

În cazul unui experiment factorial la două nivele pentru trei factori, coeficientul $\mathbf{b_1}$ se calculează astfel:

$$b_1 = \begin{Vmatrix} +1 \\ -1 \\ +1 \\ -1 \\ +1 \\ -1 \\ +1 \\ -1 \end{Vmatrix} \cdot \begin{Vmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{Vmatrix} = \frac{(y_1 + y_3 + y_5 + y_7) + (y_2 + y_4 + y_6 + y_8)}{8}$$

(5.25)

În mod similar se calculează şi ceilalți coeficienți: $\mathbf{b_0, b_2}$ şi $\mathbf{b_3}$. Modelul experimentului factorial de tipul n=$\mathbf{2^3}$ care ia în considerare şi interacțiunea factorilor are următoarea expresie:

$$\mathbf{y^n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3}$$

https://biblioteca-digitala.ro / https://unibuc.ro

(5.26)

pentru estimarea coeficienţilor $\beta_{12},\beta_{13},\beta_{23}$ (efectul interacţiunii duble) şi β_{123} (efectul interacţiunii triple) nu sunt necesare experimente suplimentare. Valorile de estimare $b_{12},b_{23},b_{13},b_{123}$ pentru coeficienţii care evaluează interacţiunile sunt determinate în acelaşi mod ca şi coeficinţii termenilor liniari. De exemplu, coeficientul b_{12} se calculează în felul următor:

$$b_{12} = \begin{Vmatrix} +1 \\ -1 \\ -1 \\ +1 \\ +1 \\ -1 \\ -1 \\ +1 \end{Vmatrix} \cdot \begin{Vmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{Vmatrix} = \frac{(y_1 + y_4 + y_5 + y_8) - (y_2 + y_3 + y_6 + y_7)}{8} \qquad (5.27)$$

Dispersiile pentru erorile de estimare ale coeficienţilor $b_1,b_2....b_k$ care în cazul experimentului ortogonal se determină independent, sunt egale: $\sigma^2_{b1} = \sigma^2_{b2} = ... = \sigma^2_{bk}$. Aceste dispersii au valoarea de estimaţie $s^2_{b_i} = \frac{s^2_e}{n}$, $i = 1,2,....k$ unde s^2_e reprezintă valoarea de estimare pentru fluctuaţia răspunsului evaluată din măsurări repetate. Semnificaţia coeficienţilor $b_1,b_2,....b_k$, adică semnificaţia efectului liniar al factorilor $x_1,x_2,...,x_k$ asupra răspunsului se stabileşte cu testul t (Student). Astfel se calculează pentru fiecare coeficient mărimea:

$$t_{exp} = \frac{s_i}{s_{bi}} \qquad (5.28)$$

Ipoteza semnificaţiei coeficientului b_i se acceptă pentru un nivel de semnificaţie α, dacă $t_{exp} > t_{(1-\alpha/2;\nu)}$

Concordanţa dintre polinomul evaluat şi valoarea adevărată a funcţiei de răspuns necunoscută se verifică cu ajutorul testului F:

https://biblioteca-digitala.ro / https://unibuc.ro

$$F_{exp} = \frac{s_i^2}{s_e^2} \qquad (5.29)$$

Ipoteza că polinomul evaluat se deosebeşte semnificativ de funcţia de răspuns adevărată necunoscută, adică cele două funcţii nu concordă, se acceptă dacă: $F > F(\alpha, \nu_1, \nu_2)$. Prin $F(\alpha, \nu_1, \nu_2)$ s-a notat parametrul testului F, corespunzător unui nivel de semnificaţie α, iar ν_1 şi ν_2 grade de libertate.

5.2.2.2 Experimente factoriale fracţionate

În cazul în care este suficientă aproximarea suprafeţei de răspuns printr-un polinom liniar (hiperplan), nu sunt necesare toate măsurătorile unui experiment factorial $n = 2^k$. Atunci când interacţiunile lipsesc, se recurge la fracţiuni (jumătăţi, sferturi) dintr-un experiment total, care prezintă însă proprietăţi de ortogonalitate.

Experimentele factoriale fracţionate nu sunt foarte eficiente deoarece interacţiunile dintre factori există şi sunt în general numeroase. De asemenea, experimentele factoriale fracţionate nu oferă date pentru calcularea dispersiei experimentelor. În consecinţă, experimentele factoriale fracţionate sunt indicate atunci când dispersia experimentelor este cunoscută *a priori*.

5.2.2.3. Experiment factorial la trei nivele

Zona din jurul optimului funcţiei de răspuns, zonă care se cere evaluată mult mai corect, se aproximează de obicei cu un polinom de ordinul II de forma următoare:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \qquad (5.30)$$

Sunt mai multe planuri speciale elaborate pentru evaluarea funcţiei de răspuns de ordinul II ca de exemplu: planul experimental compoziţional [5] şi planul experimental rotabil [15].

Planul experimental compoziţional este alcătuit din:

https://biblioteca-digitala.ro / https://unibuc.ro

- un experiment factorial complet la două nivele 2^k, pentru k<5 şi un experiment factorial fracţionat 2^{k-1} pentru n≥5;

- 2k puncte experimentale: $(\pm\alpha,0,0,...,0)$, $(0,\pm\alpha,0,...,0)$, ... , $(0,0,...,\pm\alpha)$ situate la distanţa α de centru.

- k_0 puncte experimentale în centrul planului.

În concluzie, pentru un experiment compoziţional numărul total de măsurători este:

$$n = \begin{cases} 2^k + 2k + k_0, & k < 5 \\ 2^{k-1} + 2k + k_0, & k \geq 5 \end{cases} \qquad (5.31)$$

În cazul experimentului compoziţional, valorile parametrului α se aleg astfel încât să se asigure ortogonalitatea rezultatelor experimentului.

Box şi Hunter [15] elaboreaza un plan experimental compoziţional rotabil pentru evaluarea suprafeţelor de răspuns de ordinul II. Pentru ca experimentul compoziţional să fie rotabil se alege parametrul α astfel încât să satisfacă condiţia de rotabilitate $\alpha = 2^{r/4}$, und r este exponentul tipului de experiment factorial la două nivele. De exemplu, pentru experimentul factorial fracţionat de tip 2^{k-1} parametrul α ia valoarea: $\alpha = 2^{(k-1)/4}$. Matricea de programare a unui experiment rotabil nu este ortogonală şi de aceea calcularea coeficienţilor termenilor pătratici ai polinomului de aproximare este mai complicată.

Bibliografie

1. C.Liteanu, I.Rîcă, *Optimizarea proceselor analitice*, Ed.Academiei Române 1985.
2. M.Feinberg, *L'actualite chimique* 3, 201, 1991.
3. J.S.Hunter, *Experimental Designs: Fractionals in Chemometrics* (Kowalski B., Ed.), Reidel Pub.Co., Dordrecht, 1984.
4. S.N.Deming, *Experimental Designs :Response Surfaces in Chemometrics* (Kowalski B.,Ed.), Reidel Pub.Co., Dordrecht, 1984.

https://biblioteca-digitala.ro / https://unibuc.ro

5. W.G. Cochran, G.M. Cox, *Experimental Design*, John Wiley, London, 1957.

6. J. Cea, *Optimization theory et algorythmes*, Dunod, Paris,1971.

7. E. Polak, *Computational Methods in Optimization*, Academic Press, London, 1971.

8. R.L. Fox, *Optimization Methods for Engineering Design*, Adison-Wesley Publishing Reading Mas. 1971.

9. W.J. Youden, *Statistical Methods for Chemists*, J.Wiley, N.York, 1962.

10. W. Spendley, G.R. Hext, F.R. Himsworth, *Technometrics* 4, 441 1962

11. J.A. Nedler, R. Mead, *Comput.J.* 7, 308, 1965.

12. G.E.P. Box, K.B. Wilson, *J. R. Statist. Soc.* B, 13, 1, 1951.

13. G.E.P. Box and D.W. Behnken, *Technometrics* 2, 445, 1960.

14. N.R. Draper, H. Smith, *Applied Regression Analysis*, John Wiley, New York, 1966.

15. G.E.P. Box, J.S. Hunter, *Ann. Math. Statist.* 28, 195, 1975.

https://biblioteca-digitala.ro / https://unibuc.ro

6 Procesarea informației chimice prin metodele inteligenței artificiale

6.1 Rețele neuronale

6.1.2 Introducere

Rețelele neuronale artificiale au fost concepute pentru rezolvarea unor probleme specifice inteligenței artificiale, dintre care cele mai semnificative sunt: sistemele expert, recunoaşterea de forme (configurații), procesarea imaginii, controlul automat al sistemelor, modelarea datelor analitice, etc. Inteligența artificială a fost definită ca fiind studiul proceselor inteligenței umane prin intermediul metodelor computaționale [1]. Importanța cercetărilor asupra sistemului nervos al diferitelor organisme aflate pe variate trepte de evoluție a materiei vii este atât teoretic, prin studii fundamentale asupra proceselor gândirii, cât şi practic, prin transferul de tehnologie de la biosisteme spre tehnic şi tehnologie [2,3]. Relațiile complexe dintre organisme şi mediul ambiant, ca şi necesitatea funcționării coordonate a subsistemelor unui organism, impun circulația permanentă de informație referitoare atât la parametrii mediului extern cât şi la funcționarea subsistemelor constituente. Sistemul nervos constituie o amplă rețea de comunicație în cadrul organismului, rețea în care semnalele utilizate constau din fluxuri nervoase (impulsuri). Deşi sistemul nervos central (SNC) procesează continuu informație relevant situațiilor naturale mai rapid decât orice sistem cibernetic, experiențele de psihofiziologie uman arată că intervalul de timp scurs din momentul recepționării unui stimul şi până în momentul declanşării răspunsului nu poate coborî sub 100 ms, reacție care poate fi suplimentar întârziată de factori specifici ființelor vii. Sistemele vii sunt superioare sistemelor cibernetice, cu timpi specifici de răspuns sub 1 ms, în cel puțin următoarele privințe [4,5]:

- *capacitate de analiză şi sinteză,*
- *flexibilitate,*
- *fiabilitate,*
- *redundanță,*

https://biblioteca-digitala.ro / https://unibuc.ro

- *miniaturizare*,

- *grad înalt de conectivitate*.

Rezultă posibilitatea rafinării implementării proceselor de calcul prin revizuirea atât a arhitecturii sistemelor artificiale, cât şi la nivel software.

Creierul uman, într-un volum de circa 1,5 dm^3 conţine între 10^{10} şi 10^{12} neuroni care consumă o putere totală de câteva zecimi de watt. Experienţe de neurofizică sugerează că elementele neuronale ale SNC se caracterizează printr-o viteză de calcul de ordinul milisecundelor, aceasta înseamnă un volum total de aproximativ 100 de calcule în intervalul de reacţie. Cele mai recente şi sofisticate programe de inteligenţă artificială (AI) cu destinaţii comparabile, nu sunt atât de generale şi necesită milioane de calcule. Soluţionarea rapidă a unor probleme computaţionale complexe este fezabilă pe două căi. Prima constă în transmiterea cu o viteză mult mai mare a informaţiei între elementele de procesare. Sistemele vii sunt însă extrem de lente, capacitatea de transmisie axonală fiind cuprinsă intre 50 şi 100 bit/s, iar sistemele tehnice, deşi caracterizate de o capacitate de transmisie pe canal remarcabilă, au la dispoziţie numai circuite integrate monolitice semiconductoare (IC) cu fan-out-ul (< 20) incomparabil cu cel al neuronilor din SNC (circa 10^4). A doua cale presupune ipoteza că structurile complexe de rezolvare a problemelor sunt codificate global, în configuraţia interconexiunilor unităţilor de procesare, astfel încât nu este necesar circulaţia unei cantităţi mari de informaţie. Acest mod de abordare a condus la conceptul de *conexionism*.

Calculatoarele secvenţiale clasice (maşini *von Neumann*) nu sunt concepute pentru a opera asupra unui număr mare de variabile care se influenţează reciproc şi, în plus, solicită date perfect definite. Reţelele neuronale formale, denumite aici alternativ modele cu *procesare distribuită paralel* (*modele PDP*), *modele conexioniste* sau *modele neuromorfice*, sunt concepute pentru rezolvarea unor asemenea probleme specifice *inteligenţei artificiale*, dintre care cele mai semnificative sunt: *sistemele expert, recunoaşterea de forme* (*configuraţii*) *procesarea imaginii, controlul automat al sistemelor*, etc.

https://biblioteca-digitala.ro / https://unibuc.ro

In figura 6.1 este prezentată schematic o rețea neuronală artificială .

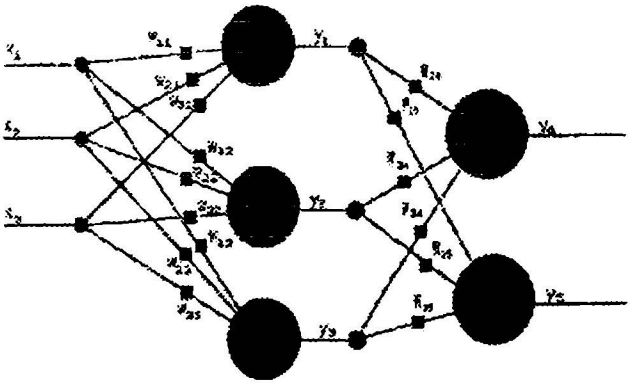


Figura 6.1. *Schema unei rețele neuronale cu un strat ascuns (intermediar).*

Rețeaua neuronală are capacitatea de a înmagazina informații diverse şi de a le face apoi disponibile pentru utilizare. Asemănarea cu creierul provine din faptul că informațiile sunt înmagazinate de rețea printr-un proces de învățare şi aceste informații sunt depozitate în conexiunile interneuronale cunoscute drept ponderi sinaptice şi nu în unitățile de procesare (neuroni). In general rețelele neuronale sunt caracterizate de trei elemente şi anume: modelul neuronului, arhitectura rețelei şi algoritmul de învațare folosit.

6.1.3 Noțiuni de teoria modelării

Modelarea este o metodă de cercetare a unor sisteme sau procese, numite originale, prin studiul asupra altor sisteme mai simple ca structură şi proprietăți care se află în anumite raporturi cu originalele. Modelele constituie imagini homomorfe ale sistemelor originale, care păstrează proprietățile esențiale ale acestora în sensul unor aplicații specifice. Parametrii unui model dat pot fi modificați pe parcursul modelării, crescând astfel eficiența sa. Modificarea modelelor se poate face extern, adoptând o structură axiomatică diferită, sau intern, prin evidențierea unor relații noi în interiotul modelului existent. Neuronul constituie elementul de bază morfologic şi fiziologic al

https://biblioteca-digitala.ro / https://unibuc.ro

sistemului nervos, fiind o celulă excitabilă cu un înalt grad de specializare, care asigură recepția, codificarea, transmisia și prelucrarea informației conținută de stimuli.

Un neuron biologic este alcătuit dintr-un corp neuronal (*soma*) și din două tipuri de prelungiri: *axoni* și *dendrite* (Figura 6.2). Organizarea internă a neuronului este similară cu a oricărei alte celule.*Corpul neuronului* (*pericarion*) poate avea dimensiuni foarte diferite, de la $5 \div 10$ μm (neuronii mici din scoarța cerebelului) până la $120 \div 125$ μm (neuronii mari din măduva spinării). El posedă un *nucleu* în jurul căruia este dispusă *citoplasma neuronală* (*neuropasma*) care conține *organite celulare comune*, *neurofibrile* și *incluziuni citoplasmatice*. Dintre organitele celulare comune sunt de menționat: (*i*) *corpusculii Nissl*, (*ii*) *ribozomii* formați din acizi nucleici (ADN și ARN) cu rol în sinteza proteinelor, (*iii*) *lizozomii*, care conțin enzime (hidrolaze) implicate în procesul de digestie celulară, și (*iv*) *mitocondrii* responsabile de respirația celulară. Citoplasma neuronală conține și organite specifice, în special *neurofibrile* formate din proteine dispuse în spirală.

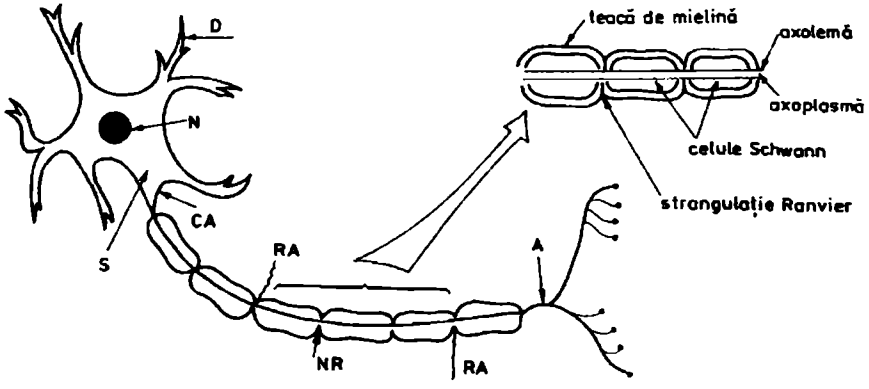


Figura 6.2
Structura schematică a unui neuron biologic;
A axon, FA fibre axonale, RA ramuri axonale inserate în dreptul nodurilor Ranvier, CA canal axonal.

Citoplasma neuronală se află într-o strânsă dependență de *nucleul* celulei nervoase, care este constituit dintr-o veziculă cu aspect clar ce conține unul sau mai mulți *nucleoli*. Acizii nucleici constituie componenții specifici ai nucleului (ADN este

https://biblioteca-digitala.ro / https://unibuc.ro

localizat în nucleoli). Învelişul nucleului este format dintr-o membrană dublă, cu pori, care face posibil transportul de la nucleu la citoplasmă a acizilor ribonucleici mesageri (purtători ai mesajelor privitoare la sinteza proteinelor).

Dendritele sunt prelungiri citoplasmatice în general bogat ramificate; ele sunt separate de mediu printr-o membrană prelungire a membranei plasmatice a corpului neuronal. Citoplasma dendritelor conţine corpusculi Nissl, mitocondrii şi neurofibrile. Dendritele primesc excitaţii şi conduc impulsurile nervoase spre pericarion, deci în sens aferent.

Axonul, prelungire citoplasmatică unică pentru fiecare celulă de lungime mult mai mare decât dendritele, conduce impulsurile nervoase în sens eferent. Diametrul axonilor variază de la 1 μm la circa 20 μm, iar lungimea sa este cuprinsă între 10 μm (axonii neuronilor din SNC) la mai mult de 1 m (axonii neuronilor motori din coarnele anterioare ale măduvei spinării). Elementele constitutive principale ale axonului sunt redate schematic în figura 6.2.

Se observă din figura 6.3 că expansiunea axonică este formată din *axoplasmă* conţinând neurofibrile, lizozomi şi mitocondrii, precum şi dintr-o serie de învelişuri. Primul dintre acestea este *axolema* (membrana axonului) care este implicată major în procesele de excitaţie şi în cele de transmitere a impulsurilor nervoase. Al doilea înveliş este *teaca de mielină*, întreruptă la intervale neregulate de $0{,}1 \div 2$ mm de nodurile (strangulaţiile) Ranvier. Deoarece nu conduce electric, mielina joacă un rol important în propagarea impulsului nervos de-a lungul fibrei. Astfel, la fibrele mielinizate în transmiterea semnalului sunt implicate numai porţiunile de membrană liberă (din nodurile Ranvier), regiuni care permit schimburi ionice cu mediul (figura 6.3). Ca urmare, transmisia se face cu pierderi reduse de energie şi, ceea ce este mai important, cu perturbaţii mai mici. În general, din axon pleacă o ramură laterală ce se inserează la nivelul unui nod Ranvier şi care se îndreaptă spre alţi neuroni, fie direct, fie prin intermediul altor ramuri axonale. Aceste fibre colaterale au rol în mecanismele de reverberaţie sau de inhibiţie prin retroacţiune (*feedback*) sau prin anticipaţie (*feedforward*).

https://biblioteca-digitala.ro / https://unibuc.ro

Deşi impulsul nervos se poate propaga în ambele sensuri de-a lungul unui axon, totuşi *in vivo* propagarea se face unidirecţional şi anume *ortodromic*: de la soma neuronală spre terminaţiile axonale. Propagarea unidirecţională a impulsului nervos se datorează unor contacte anatomice interneuronale numite *sinapse*, care se comportă ca nişte diode lăsând să treacă potenţialele de acţiune numai într-un sens, iar în sens opus prezentând o impedanţă foarte mare.

Chiar dacă din punct de vedere morfologic există o întreagă varietate de sinapse, totuşi din punct de vedere funcţional se pot diferenţia în esenţă două tipuri :

1. Sinapse cu transmisie chimică (figura 6.3a);
2. Sinapse cu transmisie electrică prin curenţi locali (figura 6.3b).

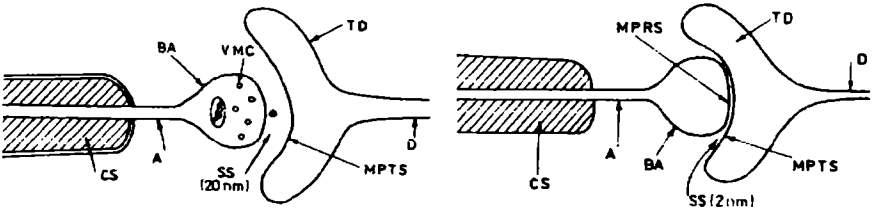
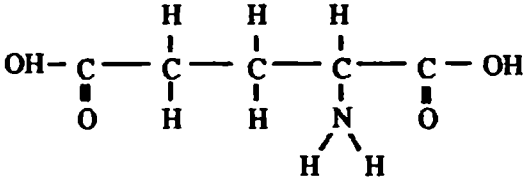


Figura 6.3
Reprezentarea schematică a sinapselor neuro-dendritice;
a) cu transmisie chimică, b) cu transmisie electrică.

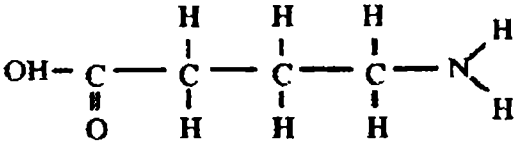
În cazul contactelor directe dintre celulele piramidale ale cortexului, mediatorul chimic în transmisia sinaptică excitatorie este glutamatul



Glutamat

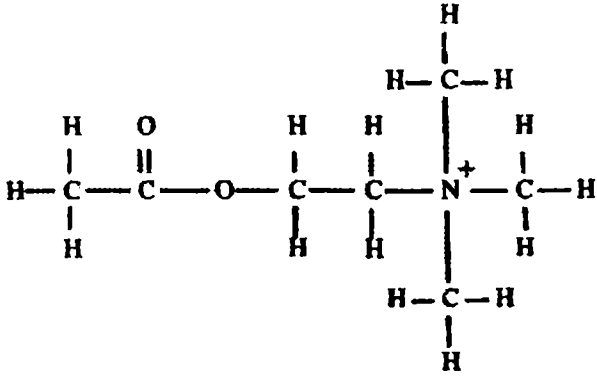
Neurotransmiţătorul pentru sinapsele inhibitorii este acidul gama-aminobutiric (GABA)

https://biblioteca-digitala.ro / https://unibuc.ro



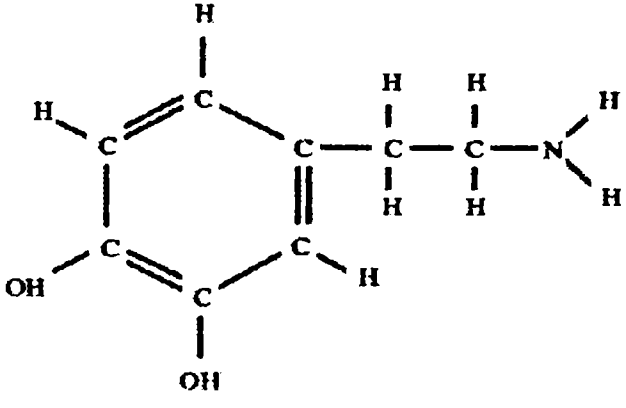
GABA

Hipocampusul este controlat de *septum* şi implicit de programele fundamentale (configuraţii de activitate) ale hipotalamusului. Neurotransmiţătorul pentru septum este acetilcolina (acelaşi mediator ca şi în cazul joncţiunilor neuromusculare)



Acetilcolina

Activităţi intense şi de durată ale formaţiunilor nigrostriate stimulează secreţia unui neurotransmiţător specific, dopamina, care blochează efectele produse de *globus pallidus* permiţând excitarea simultană a numeroase regiuni restrânse ale cortexului motor.



Dopamina

https://biblioteca-digitala.ro / https://unibuc.ro

Conexiunile interneuronale se realizeaz prin contactele sinaptice, specificitatea reţelelor neuronale fiind determinate in esenţă, nu atât de numrul şi natura neuronilor constituienţi, cât de numrul extrem de mare al interconexiunilor şi caracterul tridimensional al reţelei.

Modelarea şi simularea reţelelor neuronale sunt importante atât pentru elaborarea de modele de procesare a informaţiei în sistemul nervos animal (*"computational neuroscience"*), cât şi pentru conceptele de origine neuronal operante în sistemele inteligente artificiale (*"neural computing"*). Problema central în modelare este alegerea unui *model optimal*, care implică existenţa unor criterii de comparaţie.

In figura 6.4 este prezentat modelul general al unui neuron formal. Astfel rolul neuronului este de a primi semnale aferente $x = (x_1, x_2, \ldots x_n)$ de la alţi neuroni (presinaptici) pentru a furniza (procesa sau calcula) un semnal de ieşire y neuronilor formali postsinaptici sau la ieşirea reţelei neuronale.

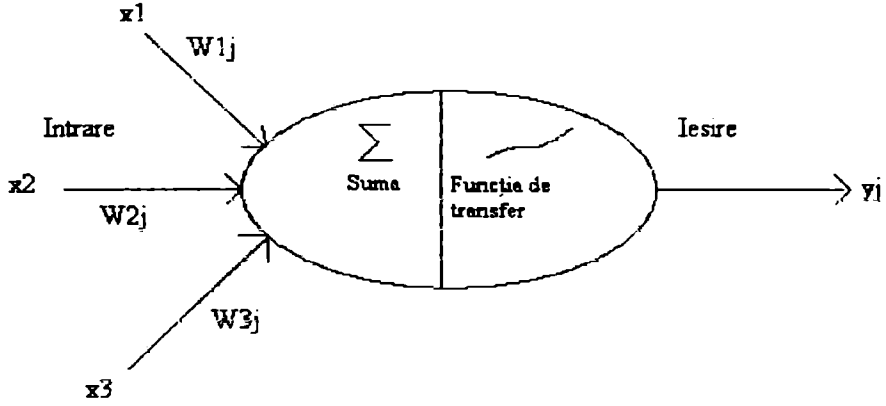


***Figura 6.2** Modelul general al unui neuron formal*

In figura 6.2, x reprezintă intrările, w reprezintă ponderile sinaptice, iar f funcţia de transfer (activare). Unitatea de procesare elementară însumează intrările ponderate iar suma rezultată constituie argumentul unei funcţii neliniare de transfer care include valoarea de prag θ. Pentru cei mai mulţi dintre neuroni, funcţia de transfer (activare) este sigmoidală (reprezentarea sa grafică este o curbă de forma literei S). Dacă notăm

https://biblioteca-digitala.ro / https://unibuc.ro

cu I suma ponderată a semnalelor de intrare, atunci expresia funcției de transfer este următoarea:

$$f(I) = \frac{1}{1 + e^{-I}}$$

Important este ca această funcție să fie monoton crescătoare şi mărginită la ambele capete. O altă formă preferată în aplicații a funcției de activare este tangenta hiperbolică deoarece este uşor inversabilă iar inversa sa este simetrică față de origine.

Rețelele neuronale formale pot fi împărțite în două clase pe baza măsurii în care utilizează modelul procesării distribuite [6]:

- *modele locale*, în care activitatea unei singure unități de procesare reprezintă gradul de participare a unei unități conceptuale cunoscute;

- *modele distribuite*, în care o configurație a activităților unui numr mare de unități determină acelasi grad de participare.

Ca exemple, modelul activării interactive este un model PDP local, în timp ce rețeaua Hopfield este un model PDP distribuit.

Referitor la rețelele neuronale, se utilizează frecvent termenul de *paradigmă*. În acest context, paradigma cuprinde tipic descrierea formei şi funcției unității elementare de procesare (celul, nod, neuron,...), topologia rețelei, care specifică configurația interconexiunilor dintre unități, precum şi o regulă de adaptare (învățare, instruire) a ponderilor interconexiunilor. Trei clase de sisteme pot constitui paradigme ale sistemelor inteligente: *sistemele continui*, *sistemele logice* şi *sistemele simbolice*. Acestora le corespund în calculatoarele convenționale respectiv nivelele de *circuististică*, de *logică* şi de *programare*. Paradigmele rețelelor neuronale, deşi distincte în detaliu, selectează o submulțime comună de atribute tipice rețelelor neuronale biologice [7]:

- *elemente simple de procesare,*
- *nivel ridicat de conectivitate,*
- *procesare distribuită paralel a datelor,*
- *funcție de transfer neliniară pentru neuroni,*
- *retroacțiune ("feed-back"),*

https://biblioteca-digitala.ro / https://unibuc.ro

- *propagare neamortizată şi cu viteză constantă energiei,*
- *procesare nonalgoritmică,*
- *adaptabilitate,*
- *autoorganizare,*
- *toleranţă la defecte,*
- *răspunsuri relevante la date de intrare vagi ("fuzzy"),*
- *. putere de generalizare.*

Abilitatea procesării simbolice către care tind modelele performante de reţele neuronale actuale este o consecinţă tehnică nebănuită a calculatoarelor digitale, care nu au fost iniţial concepute în acest scop. În mod similar poate fi privit facilitatea de a raţiona logic şi rezolva probleme matematice, care apare ca o "deficienţă tehnică" iniţial neaşteptată a creierului uman, în mod indiscutabil inutil supravieţuirii în zorii apariţiei speciei umane [8÷10].

Problema centrală în modelarea reţelelor neuronale este alegerea unui *model optimal*, care implică existenţa unor criterii de comparaţie. Aceste criterii sunt mai mult sau mai puţin semnificative în contextul unor aplicaţii particulare. Alegerea unui model existent sau construcţia unui model nou, depind de mai mulţi factori cu ponderi diferite:

- disponibilitatea şi tipul facilităţilor computaţionale de prelucrare a datelor şi simulare asistate de calculator;
- suport tehnic şi financiar;
- experienţa şi nivelul ştiinţific al grupului de cercetare;
- antecedente istorice, etc.

Sursele de incertitudine (nedeterminare), care afectează modelarea sunt clasificate în trei categorii distincte, având la origine următoarele [11] :

https://biblioteca-digitala.ro / https://unibuc.ro

1. măsurători imprecise sau nesigure (cu o anumită lipsa de acuratețe);
2. abateri aleatorii;
3. descrieri sau prezentări vagi (incerte).

Se pot considera că aceste trei clase calitativ diferite de la care poate proveni incertitudinea în procesul modelării, conduc la trei tipuri de modele distincte [2, 12, 13]:

1. *modelul determinist;*
2. *modelul probabilistic (stohastic);*
3. *modelul vag (fuzzy).*

Modelul matematic al unui sistem original sau proces fizic este constituit din ansamblul supozițiilor şi ecuațiilor matematice care descriu complet starea şi evoluția subiectului modelării. *Modelul fizic* al unui sistem original sau proces fizic reprezintă realizarea materială concretă (tehnică) a modelului matematic ce constituie echivalentul conceptual al modelului fizic şi pe care îl include intrinsec.

Modelarea matematică poate fi întreprinsă în cadrul unuia dintre cele trei moduri de tratare a incertitudinii prezentate mai sus.

Modelul determinist presupune că atât parametrii sistemului cât şi variabilele de intrare sunt mărimi deterministe, aşadar mărimi a caror evoluție în timp poate fi anticipată (determinată) dacă se cunoaşte evoluția lor anterioară şi condițiile inițiale. Acest model corespunde descrierii comportamentului rețelelor paucineuronale, cu un număr relative redus de neuroni.

Modelul probabilist presupune că atât parametrii sistemului cât şi variabilele de intrare sunt mărimi stohastice. În general, acuratețea şi utilitatea modelelor probabiliste sunt exclusiv dependente de valorile asociate acestor mărimi statistice şi a funcțiilor de distribuție asociate a priori. Imprecizia provine astfel din incertitudinea intrinsecă a rezultatului final; odată efectuată însă experiența, acest rezultat este perfect cunoscut.

Modelul vag presupune vagi atât parametrii sistemului cât şi variabilele de intrare ale sistemului original, pornind de la imprecizia datelor experimentale şi încheiind cu imprecizia descriptivă a termenilor lingvistici. Variatele tipuri de

https://biblioteca-digitala.ro / https://unibuc.ro

nedeterminare pot fi specificate în cadrul teoriei mulțimilor vagi, astfel încât procesarea datelor se poate desfășura independent de natura investigațiilor. În comparație cu modelele statistice la care mulțimea rezultatelor posibile este calculată pornind de la o funcție de distribuție probabilistă acceptată iar incertitudinea asupra rezultatului dispare după desfășurarea experimentului, la modelele vagi, rezultatul însuși rămâne vag dar oferă maximum de informație în context. Conceptele teoriei mulțimilor vagi sunt transpuse în tehnici de raționament aproximativ cu aplicații în cele mai diverse direcții ale AI. Raționamentele aproximative uni- și multidimensionale oferă posibilitatea de a deduce proprietăți relevante necunoscute din informația disponibilă a unei baze de date [12].

Modelele fizice (tehnice) neuronale se bazează pe fenomene electromecanice, termice, magnetice, electronice, etc., fiind cunoscute în general sub denumirea de neuristori. Așadar neuristorii sunt dispozitive active care simulează un set de caracteristici funcționale neuronale, care pot varia de la o aplicație la alta, păstrând însă un set comun de atribute specifice neuronului.

6.1.4 Modele de rețele neuronale

6.1.4.1 Rețele neuronale formale cu instruire supervizată (memorii asociative)

Conceptul de *memorie asociativă* caracterizează abilitatea unui sistem dinamic de a parcurge o secvență de reprezentări interne sau de reconstituire a unei reprezentări complexe pornind de la o versiune distorsionată sau incompletă a sa. Memoria asociativă are la bază conceptul de reprezentare distribuită a informației prin stări colective ale rețelei neuronale. Dinamica interacțiilor poate fi făcută să corespundă structurii informaționale, fără un echivalent fizic direct în sistem, realizându-se prin efecte globale și reflectându-se în procesele de redare a informației memorate. Problemele fundamentale ridicate de memoria asociativă sunt: *diafonia* (*"cross-talk"*), *invarianța* și *inabilitatea* de a opera cu structuri.

https://biblioteca-digitala.ro / https://unibuc.ro

Structurile interacţiilor sunt puse în corespondenţă cu structurile cunoaşterii; chiar dacă acestea din urmă nu au un echivalent fizic direct in sistem, se pun în evidenţă prin efectele stărilor colective şi se reflectă în procesele de *reproducere*, *redare* sau *reconstituire* (*"recall"*) din memorie. Modul de reprezentare distribuită a informaţiei prin stări colective ale neuronilor reţelei este numit *holografic* sau *hologic*, dar eleganţa sa formală este estompată de implementarea pe calculatoare secvenţiale convenţionale.

Ca exemple de modele de reţele neuronale formale cu instruire supervizată pot fi menţionate : modelul *Sutton-Barto*, modelul *heterostat*, modelul *Hopfield-Tank*, modelul *Hoppenstead*, modelul *Hamming*, modelul *Hopfield stohastic* şi *perceptronul*.

Modelul Sutton-Barto este reprezentativ în sensul influenţei reciproce de condiţionare dintre configuraţia aşteptată şi configuraţia stimulului, care este prea complexă pentru a fi exprimată printr-o regulă de corelaţie simplă de tip hebbian; în timp ce regula lui Hebb detectează corelaţii între semnalele de intrare şi ieşire, algoritmul Sutton-Barto detectează corelaţii între "urmele" stimulilor de intrare şi modificările produse la ieşire.

Modelul heterostat introdus de Klopf şi bazat pe analogia dintre sistemele sociale şi reţelele neuronale, constituie un alt model clasic de memorie asociativă. Neuronii sunt presupuşi înzestraţi cu o caracteristică hedonistică, conform căreia activitatea neuronală se desfăşoară în sensul obţinerii excitaţiei şi evitării inhibiţiei.

Modelul Hopfield-Tank, bazat pe determinarea condiţiilor de detonare a neuronilor, utilizează tehnici de relaxare pentru satisfacerea constrângerilor şi *întărire simulată* (*"simulated annealing"*) cu scopul de a evita fixarea în minime locale. In mod special se justifică aplicaţiile reţelei Hopfield-Tank ca memorie asociativă şi ca rezolvant în problemele de optimizare.

Modelul Hoppenstead, bazat esenţialmente pe sincronizarea frecvenţelor de detonare a neuronilor, poate fi analizat prin asocierea unei funcţii de energie cu *calare pe fază* (*"phase-locking"*) a unei reţele de oscilatori electrici şi conduce la conceptul de *oscilator neuronal controlat de tensiune*. Ultimile două modele prezentate constituie exemple reprezentative de sisteme PDP clasice.

https://biblioteca-digitala.ro / https://unibuc.ro

Modelul Hamming implementează calculul distanței Hamming, care măsoară diferența a două configurații. In cazul procesării informației codificate binar, distanța Hamming este definită prin numărul de biți diferiți între cele două configurații comparate, având valoarea maximă egală cu numărul biților configuratiei de comparat (disjuncție) și valoarea minimă egală cu zero (identitate).

Rețeaua Hamming este superioară în raport cu rețeaua Hopfield in recunoaşterea de configurații, în sensul că necesită mai puține conexiuni decât o rețea Hopfield similară. Modelele Hopfield şi Hamming fac parte din clasa rețelelor neuronale cu instruire supervizată, adică învățarea se face într-o primă etapă în care sunt prezentate la intrare configurațiile de referință şi are ca efect adaptarea corespunzătoare a ponderilor interconexiunilor. Ulterior, în etapa a doua, o configurație necunoscută este prezentată la intrare şi este asociată cu cel mai "apropiat" exemplar din setul memorat, obținându-se fie exemplarul restaurat al configurației de intrare (rețeaua Hopfield), fie clasa corespunzătoare configurației de la intrare (rețeaua Hamming).

Perceptronul constituie o clasă reprezentativă de rețele neuronale introdusă de Rosenblat [14], care este capabilă de instruire supervizată, având astfel caracterul unei *memorii asociative.* Caracterul paralel al procesării informației este finalizat în modelele matematice ale *suprafețelor decizionale (modelul liniar şi modelul descreşterii gradientului),* care susțin premiza fundamentală a lui Rosenblatt relativă la caracterul probabilistic al perceptronului. Dacă teorema de convergență a perceptronului oferă a priori succesul algoritmului recursiv de instruire, în cazul perceptronilor multistratificați pentru care nu există asemenea teoreme de convergență, funcționează criterii particulare de suficiență (*"credit asignment problem"*).

Modelele matematice fundamentale ale rețelelor neuronale includ inevitabil, modelul neuronului binar clasic, de inspirație biologică introdus de McCulloch şi Pitts [15]. Totuşi, trebuie remarcat caracterul analogic (liniar) al comportamentului unui neuron biologic: funcționarea sa se bazează pe curenți ionici care variază continuu şi mai puțin pe valori discrete cărora să li se poată asocia nivelele de *0* şi *1*. Digitizarea

https://biblioteca-digitala.ro / https://unibuc.ro

impune precizie la prelucrarea informației în sistemele biofizice esențialmente imprecise și totuși sistemele neuronale biologice sunt procesoare cu o excelentă eficiență. Pe de altă parte, comportamentul predominant determinist al rețelelor paucineuronale, pe măsura creșterii complexității rețelei, devine probabilistic în cazul rețelelor plurineuronale. O explicație calitativă posibilă o constituie activitatea rețelelor neuronale în consens cu legile fundamentale ale naturii (relații statistice) și mai puțin împotriva lor (relații deterministe). In acest sens este relevantă comparația energetică între cel mai perfecționat microprocesor actual, care consumă aproximativ 10^{-9} J per operație, în timp ce neuronii biologici utilizează numai 10^{-16} J pentru același proces. Mai mult, în sistemele digitale datele și procesele computaționale trebuiesc convertite în cod binar, proces care necesită aproximativ 10.000 modificari digitale de tensiune per operație. Sistemele analogice realizează aceeași operație într-un singur pas, micșorând astfel substanțial timpul dar mai ales consumul de energie cu un factor de 10.000. Se consideră că cea mai reprezentativă caracteristică a circuitelor neuronale analogice la funcționarea în medii cu parametri imprevizibili o constituie *abilitatea lor de a reprezenta informația in context*. Rețelele neuronale răspund la diferențele în amplitudine ale semnalelor și mai putin la valoarea lor absolută, eliminând astfel necesitatea unei calibrări precise. Contextul poate fi constituit de o colecție complexă de configurații neuronale de referință acumulate în procesul de instruire. *Jocul dintre context și adaptabilitate este principiul fundamental al paradigmei rețelelor neuronale*,care impune anumite restricții specifice. Deoarece numai modificările și diferențele transmit informație, o permanentă variație a intensității stimulilor este o necesitate pentru sistemele neuronale și nu un impediment, ca în cazul sistemelor digitale. Această caracteristică definitorie a sistemelor vii, care impune variația stimulilor pentru o percepție optimă, plasează clar sistemele neuronale în interiorul lumii observabile, în contrast cu sistemele digitale (numerice) al căror design presupune implicit separarea dintre sistem și lumea exterioară [2].

Considerațiile de mai sus sunt cel puțin în parte susținute de realizarea relativ recentă pe suport semiconductor a retinei artificiale [16] sub forma unui circuit

https://biblioteca-digitala.ro / https://unibuc.ro

semiconductor monolitic VLSI în tehnologie CMOS. In consecință, o pondere însemnată este acordata *modelelor statistice* și *modelelor spin glass*, dezvoltate prin asociație cu proprietățile statistice ale sistemelor magnetice, în care interacțiile dintre momentele magnetice se află "în conflict" reciproc datorită unei dezordini structurale impusă rețelei.

Rețele Hopfield stohastice cu straturi neuronale intermediare (ascunse), urmăresc posibilitatea realizării de memorii asociative capabile să reconstituie configurațiile de intrare prin utilizarea parțială a rețelei. Asemenea rețele constituie mașini de calcul bazate pe reguli computaționale statistice numite *mașini Boltzmann*, prin opoziție cu calculatoarele digitale secvențiale clasice numite *mașini van Neumann*.

6.1.4.2 Rețele neuronale formale cu instruire nesupervizată (clasificatori vectoriali)

Rețelele neuronale cu instruire nesupervizată și autosupervizată sunt denumite generic *clasificatori vectoriali*. În cazul rețelelor neuronale cu instruire nesupervizată numai o singură ieșire este activă la un moment de timp. Aceste rețele organizează datele de intrare sub formă de clase sau de categorii. Rețelele neuronale cu instruire nesupervizată au nevoie de un număr de celule de ieșire egal cu numărul de categorii care se dorește memorat. De exemplu N unități de ieșire pot clasifica numai N categorii și nu 2^N ca în cazul memoriilor asociative. Prezentăm mai jos câteva exemple de rețele neuronale formale cu instruire nesupervizată.

Rețeaua Carpenter-Grosberg este reprezentativ prin comparație cu rețeaua Hamming; în sensul că asigură clasificarea exemplarelor pe măsura prezentării lor la intrare prin identificarea unor clase de configurații asemănătoare specificate prin câte un exemplar reprezentativ.

Modelul instruirii competitive, model de învățare nesupervizat corespunzător detectorului de regularitate, este în esență un algoritm de instruire statistic neasociativ și este analizat în legătură cu abilitatea rețelelor neuronale adaptive de a descoperi caracteristici relevante ale stimulilor de intrare. Cele mai reprezentative arhitecturi

https://biblioteca-digitala.ro / https://unibuc.ro

structurale şi mecanisme de instruire competitiv sunt: *modelul Linsker* şi *modelele Fukushima.*

Modelul Linsker este inspirat din proprietăţile celulelor primelor etape de procesare a informaţiei în sistemul vizual la mamifere care, dezvoltându-se la unele animale înainte de naştere, nu pot constitui rezultatul unei experienţe structurate. In acest context, sunt consacrate *metoda analizei componentei principale* pentru identificarea structurii de interes neanticipate în seturi mari de date (de exemplu *partiţionarea*) şi *principiul infomax* de prezervare la ieşire a informaţiei maxime (în sensul lui Shannon) relativ la activitatea de intrare supusă la constrângeri.

Modelele Fukushima, cognitronul şi ulterior, neocognitronul, care constituie variante perfecţionate de instruire a reţelelor neuronale multistratificate, sunt dezvoltate în principal ca răspuns la observaţia că reţelele neuronale din centrii vizuali ai creierului evoluează în direcţia recunoaşterii caracteristicilor configuraţiilor prezentate la intrare.

Modelul activării interactive este calitativ un algoritm capabil să explice elementele fundamentale ale percepţiei vizuale şi auditive.

Teoria rezonanţei adaptive continuă seria modelelor de instruire competitiv, care încearc să modeleze stabilitatea algoritmilor competitivi în cazul reţelelor plurineuronale.

Un alt grup important de modele caracterizate prin instruire autosupervizată cuprinde *algoritmii dezvoltaţi pe baza principiului darwinist de evoluţie a speciilor prin selecţie*, reflectat în reţelele neuronale prin competiţia dintre neuroni în procesele de învăţare. Asfel următoarele modele pot fi considerate cele reprezentative [2]:

Modelul Malsburg, care modifică uşor principiul evoluţionist al selecţiei printr-o combinaţie înteligentă de inhibiţie şi excitaţie locală, explicând astfel structura cortexului vizual la pisică.

Modelul Kohonen, care propune un mecanism de autoorganizare a reţelelor neuronale prin care numărul neuronilor alocaţi unei regiuni din spaţiul stimulilor de intrare este proporţional cu nivelul ei de activitate.

https://biblioteca-digitala.ro / https://unibuc.ro

Modele "genetice" de instruire, care se bazează pe un ansamblu de opțiuni inițiale (populația de cromozomi) în procesele iterative de instruire.

Modelul Hecht-Nielsen, care combină regula lui Hebb cu conceptele instruirii competitive, în scopul identificării aspectelor structurale caracteristice câmpului datelor de intrare.

Multitudinea modelelor teoretice cuprinse în literatura de specialitate şi care reflectă, în esență, intuițiile autorilor asupra funcționării creierului, au fost selecționate pentru prezentare în bună măsură subiectiv, dar ordonate după două criterii considerate de noi fundamentale. Primul corespunde clasificării propuse de Rumelhart [17], având la bază caracteristicile perechii intrare-ieşire, criteriu conform căruia se deosebesc patru clase:

1. *Autoasociatorul*,
2. *Asociatorul de configurații*,
3. *Clasificatorul de configurații*,
4. *Detectorul de regularitate*.

Al doilea criteriu, corelat însă cu primul, se bazează pe natura asistării procesului de învățare a rețelei, conform căruia se disting următoarele modalități:

- *instruire supervizată*,
- *instruire nesupervizată*,
- *instruire autosupervizată*.

6.1.5 Algoritmi de instruire a rețelelor neuronale formale

6.1.5.1 Algoritmi de instruire generalizați

a) Algoritmul instruirii exhaustive (*Heavy Learning Algorithm*)

Rețelele neuronale multistratificate pot fi caracterizate alternativ de:

https://biblioteca-digitala.ro / https://unibuc.ro

$U = \mathrm{L}(Z_1, Z_2)$

$x \longrightarrow y = y_1 + y_2,$

unde $y_1 = c(x),\ c : X \longrightarrow Y$

$y_2 = \Phi(x)\mathrm{W}\Psi(x),\ \Psi : X \longrightarrow Z_1,\ \mathrm{W} \in \mathrm{L}(Z_1, Z_2),\ \Phi : X \longrightarrow \mathrm{L}(Z_2, Y)$

Algoritmul trebuie să găsească o matrice sinaptică $\mathbf{W_P}$ care să învețe setul de instruire în sensul că:

$$(\forall)p = 1,\ldots,P,\quad c(a_p) + \Phi(a_p)\mathbf{W_P}\Psi(a_p) = b_p \tag{6.1}$$

Dacă sunt satisfăcute următoarele condiții:

$\mathrm{c} : X \longrightarrow Y$ °i $\Psi : X \longrightarrow Z_1$ continue

$\Phi : X \longrightarrow \mathrm{L}(Z_2, Y)$ continua si surjectivă

$\Psi(\mathrm{a_p})$ ortonormal între ele

atunci :

$$\mathbf{W}_P = \mathbf{W}_{P-1} - \Psi(a_p) \otimes \Phi(a_p)^{+}\left(c(a_p) + \Phi(a_p)\mathbf{W}_{P-1}\Psi(a_p) - b_p\right) \tag{6.2}$$

b) Algoritmul gradientului

Fie $\mathrm{f} : \mathrm{X} \to \mathbf{R}$ o funcție pe care dorim să o minimizăm, atunci:

1. Se alege $x_n \in X, n = 0;$
2. Dacă $\mathrm{f}'(x_n) = 0,$ procesul se încheie, altfel $x_{n+1} = x_n - \delta_n \dfrac{\mathrm{f}'(x_n)}{\|\mathrm{f}'(x_n)\|};$
3. $n \longrightarrow n+1$ si se revine la pct. 2.

Se poate arăta că dacă $\mathrm{f} : \mathrm{X} \to \mathbf{R}$ este o funcție convexă diferențiabilă şi mărginită inferior, iar dacă considerăm o secvență de "paşi"

$$\delta_n > 0, \quad \lim_{\mathrm{n} \longrightarrow +\infty} \delta_n = 0, \quad \sum_{n=0}^{\infty} \delta_n = +\infty$$

atunci următoarea secvență descrescătoare de scalari este convergentă:

https://biblioteca-digitala.ro / https://unibuc.ro

$$\theta_k = \min_{n=0,\ldots,k} f(x_n) \xrightarrow[k \to +\infty]{} v = \inf f(x) \tag{6.3}$$

Rețele neuronale neliniare.

Să considerăm în continuare următoarea rețea:

$X = \mathbf{R}^n$, $Y = \mathbf{R}^m$, $W \in L(X,Y)$, $f(x,W) = g(Wx)$ unde $g : Y \longrightarrow Y$.

Se cere determinarea unei matrici sinaptice care să optimizeze rețeaua neuronală pentru un set de configurații dat:

$$\Phi(x, \mathbf{W}) = y \tag{6.4}$$

unde am notat cu **W** matricea asociată operatorului *W*. Fiind date o funcție de evaluare *E* definită pe spațiul de ieşire *Y* şi o configurație (a,b), se caută o matrice sinaptică **W**, care să minimizeze funcția:

$$\mathbf{W} \longrightarrow H(\mathbf{W}) = E(g(\mathbf{W}a) - b) \tag{6.5}$$

Pentru cazul simplu al unei rețele neuronale cu un singur strat descrisă de funcția de propagare $g : Y \longrightarrow Y$ °i $\Phi(a, \mathbf{W}) = g(\mathbf{W}a) - b$, şi având în vedere modul de evaluare a expresiei gradientului unei funcții

$y \xrightarrow{\Psi} \Psi(y) = E(g(y))$ care conduce la $g'(y)^* E'(g(y))$, avem:

$$H'(\mathbf{W}) = a \otimes g'(\mathbf{W}a)^* E'(g(\mathbf{W}a) - b) \tag{6.6}$$

având intrările egale cu:

$$\frac{\partial H}{\partial w_{ij}} = a_j \left(\sum_{k=1}^{m} \frac{\partial g_k}{\partial y_i}(\mathbf{W}a) \frac{\partial E}{\partial y_k}(g(\mathbf{W}a) - b) \right) \tag{6.7}$$

Rezultă că dacă matricea sinaptică $\overline{\mathbf{W}}$ determină minimul lui *H*, atunci ea este soluția sistemului de ecuații:

https://biblioteca-digitala.ro / https://unibuc.ro

$$a_j\left(\frac{\partial g_i}{\partial y_i}\left(\overline{\mathbf{W}}a\right)\frac{\partial E}{\partial y_i}\left(g\left(\overline{\mathbf{W}}a\right)-b\right)\right)=0,\ (\forall)i,j \tag{6.9}$$

şi regula de instruire a reţelei va fi:

$$w_{ij}^{n+1}=w_{ij}^{n}-\varepsilon_n a_j\left(\frac{\partial g_i}{\partial y_i}\left(\mathbf{W}^n a\right)\frac{\partial E}{\partial y_i}\left(g\left(\mathbf{W}^n a\right)-b\right)\right) \tag{6.10}$$

c) Algoritmul lui Newton

Să considerăm reţele neuronale cu unul, mai multe, sau o infinitate de straturi (continuă): $x\xrightarrow{\Phi}y=\Phi(x,\mathbf{W})$, fie $\left(a^j,b^j\right)_{1\le j\le J}$ o familie de configuraţii; atunci să determinăm matrici sinaptice $\overline{\mathbf{W}}$ care să înveţe aceste configuraţii, deci:

$\Phi\left(a^j,\overline{\mathbf{W}}\right)=b^j,\ (\forall)j=1,\ldots,J$ sau mai compact $\Phi\left(\overline{\mathbf{W}}\right)=\left(g\left(\overline{\mathbf{W}}a^j\right)-b^j\right)_{1<j<J}$

Algoritmul lui Newton în acest context porneşte de la:

$$\Phi'\left(\mathbf{W}^n\right)\left(\mathbf{W}^{n+1}-\mathbf{W}^n\right)=-\delta\,\Phi\left(\mathbf{W}^n\right),\ (\forall)n\ge 0. \tag{6.11}$$

unde se presupun date matricea iniţială arbitrară $\mathbf{W}^0$ şi $\delta\ge 0.5$.

Să considerăm o reţea neuronală neliniară care asociază spaţiului de intrare $X=\mathbf{R}^n$ spaţiul de ieşire $Z=\mathbf{R}^m$, prin stratul intermediar Y. Semnalelor de intrare $x\in X$, ponderate prin intermediul matricii sinaptice $\mathbf{W}$ asociat operatorului $W\in L(X,Y)$, le corespund la ieşire răspunsurile $f(x,\mathbf{W})=g(\mathbf{W}x)$, $g:Y\longrightarrow Z$; atunci dacă funcţia de transfer g este diferenţiabilă, funcţiile $g'\left(\mathbf{W}a^j\right)$, $(\forall)j=1,\ldots,J$, sunt surjective $(\forall)W\in L(X,Y)$, iar intrările a^j sunt ortogonale pentru produsul scalar l definit pe X, avem:

$$\mathbf{W}^{n+1}=\mathbf{W}^n+\delta\,La^j\otimes\sum_{j=1}^{J}\left(g'\left(\mathbf{W}^n a^j\right)\right)^{+}\cdot\left(\frac{b^j-g\left(\mathbf{W}^n a^j\right)}{l\left(a^j,a^j\right)}\right) \tag{6.12}$$

Aseastă regulă de instruire este de asemenea de tip Hebbian.

https://biblioteca-digitala.ro / https://unibuc.ro

d) Algoritmul cu propagare inversă a erorii

Aplicarea algoritmului cu propagare inversă a erorii (*backpropagation algorithm*) presupune parcurgerea a două etape. Într-o primă fază se prezintă rețelei un vector de intrare x^k și ieșirea dorită corespunzătoare pe care o vom nota cu t^k. Semnalul de intrare se propagă prin rețea și generează un răspuns. Neuronul j din câmpul de ieșire produce semnalul:

$$o_j^k = s_j^o\left(y_j^k\right) \tag{6.13}$$

Acest semnal este comparat cu răspunsul dorit t^k. Rezultă un semnal de eroare δ_{kj} care se calculează pentru fiecare neuron de ieșire în parte:

$$\delta_{kj} = \left(t_j^k - o_j^k\right)\frac{\partial S_j^o}{\partial y_j^k},\ j = 1,2,...,p \tag{6.14}$$

În continuare se calculează modificările ponderilor pentru toate conexiunile ce intră în stratul final folosind regula următoare:

$$\sigma_{qj}^o(k+1) = \sigma_{qj}^o(k) + c\delta_{kj}S_q^h\left(h_q^k\right) \tag{6.15}$$

unde $S_q^h\left(h_q^k\right)$ este ieșirea neuronului intermediar q.

Cea de a doua fază a algoritmului corespunde propagării înapoi prin rețea a semnalului de eroare. Aceasta permite calculul recursiv al erorii conform formulei

$$\delta_{kq}^h = \frac{\partial S_q^h\left(h_q^k\right)}{\partial h_q^k}\sum_j \delta_{kj}\sigma_{qj}^o \tag{6.16}$$

Calculul de mai sus se efectuează pentru fiecare neuron q din stratul intermediar. In acest mod, eroarea se propagă înapoi prin rețea. Propagarea inversă a erorii implică un calcul de aceeași complexitate cu propagarea înainte a semnalului și, deci, ea nu este excesiv de costisitoare sub aspectul consumului de timp și a necesarului de memorie solicitat. Ținând cont de aceste considerente, se poate introduce procedura de instruire a rețelelor cu un strat ascuns (figura 6.3) prezentată mai jos.

https://biblioteca-digitala.ro / https://unibuc.ro

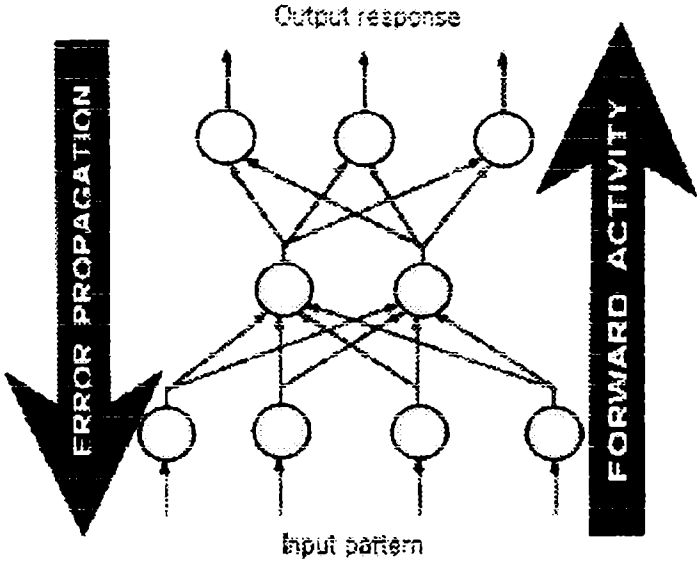


Figura 6.3 *Rețea neuronală stratificată directă cu un strat intermediar*

Algoritmul cu propagarea inversă a erorii pentru rețelele cu un singur strat intermediar

P0. Inițializare

Se inițializează ponderile rețelei cu valori mici, generate aleator. Se inițializează constanta de instruire $c \in (0,1)$.

P1. Un ciclu de instruire

Se prezintă rețelei un set de configurații de instruire (antrenare). Pentru fiecare exemplu se execută secvența *P2-P12*. In scopul evitării minimelor locale se recomandă ca ordinea de prezentare a configurațiilor de instruire să fie schimbată aleator de la un ciclu la altul.

P2. Se prezintă rețelei vectorul $\boldsymbol{x}^k = \left(x_1^k, x_2^k, ..., x_n^k\right)^T$ și vectorul ieșirilor țintă $\boldsymbol{t}^k$;

P3. Se calculează valorile de activare ale neuronilor din stratul ascuns utilizând relația: $h_q^k = \sum_{i=1}^{n} x_i^k \sigma_{ik}(q)$;

P4. Se calculează ieșirile neuronilor din stratul ascuns după expresia: $i_k^q = S_q^h\left(h_q^k\right)$;

https://biblioteca-digitala.ro / https://unibuc.ro

P5. Se calculează valorile de activare ale neuronilor din stratul de ieşire după formula: $y_j^k = \sum_q i_q^k \sigma_{ik}^o(k) = \sum_q S_q^h\left(h_q^k\right)\sigma_{iq}^o(k)$;

P6. Se calculează ieşirile reţelei cu formula: $o_j^k = S_j^o\left(y_j^k\right)$;

P7. Se calculează semnalul de eroare pentru neuronii din stratul de ieşire utilizând expresia: $\delta_{kj} = \left(t_j^k - o_j^k\right)\frac{\partial S_j^o\left(y_j^k\right)}{\partial y_j^k}$;

P8. Se calculează semnalul de eroare pentru neuronii stratului intermediar astfel: $\delta_{kq}^h = \frac{\partial S_q^h\left(h_q^k\right)}{\partial h_q^h}\sum_j \delta_{kj}\sigma_{qj}^o(k)$. Trebuie remarcat că acest semnal de eroare se calculează înainte de actualizarea ponderilor neuronilor din stratul de ieşire;

P9. Se analizează ponderile neuronilor stratului de ieşire cu relaţia: $\sigma_{qj}^o(k+1) = \sigma_{qj}^o(k) + c\delta_{kj}S_q^h\left(h_q^k\right)$, unde q este un neuron al câmpului intermediar şi j este un neuron aparţinând stratului de ieşire.

P10. Se actualizează ponderile neuronilor stratului ascuns după relaţia: $\sigma_q^h(k+1) = \sigma_q^h(k) + cx_i^k\delta_{kq}^h$, unde i este un neuron al stratului de intrare. Ordinea în care se actualizează ponderile unui nod nu este importantă;

P11. Se calculează valoarea funcţiei de eroare care este o măsură a calităţii instruirii reţelei: $E_k = \frac{1}{2}\sum_{j=1}^{p}\left(t_j^k - o_j^k\right)^2$;

P12. Iteraţia

Dacă eroarea este acceptabilă pentru toate perechile $\left\{\left(\boldsymbol{x}^k, \boldsymbol{t}^k\right)\right\}$ din mulţimea de instruire (adică dacă eroarea medie are o valoare acceptabilă), atunci învăţarea se poate considera încheiată. Altfel, vectorii de instruire sunt prezentaţi din nou reţelei în cadrul unui nou ciclu de antrenament . Diagrama schematică a procesării datelor este reprezentată în figura 6.4.

https://biblioteca-digitala.ro / https://unibuc.ro

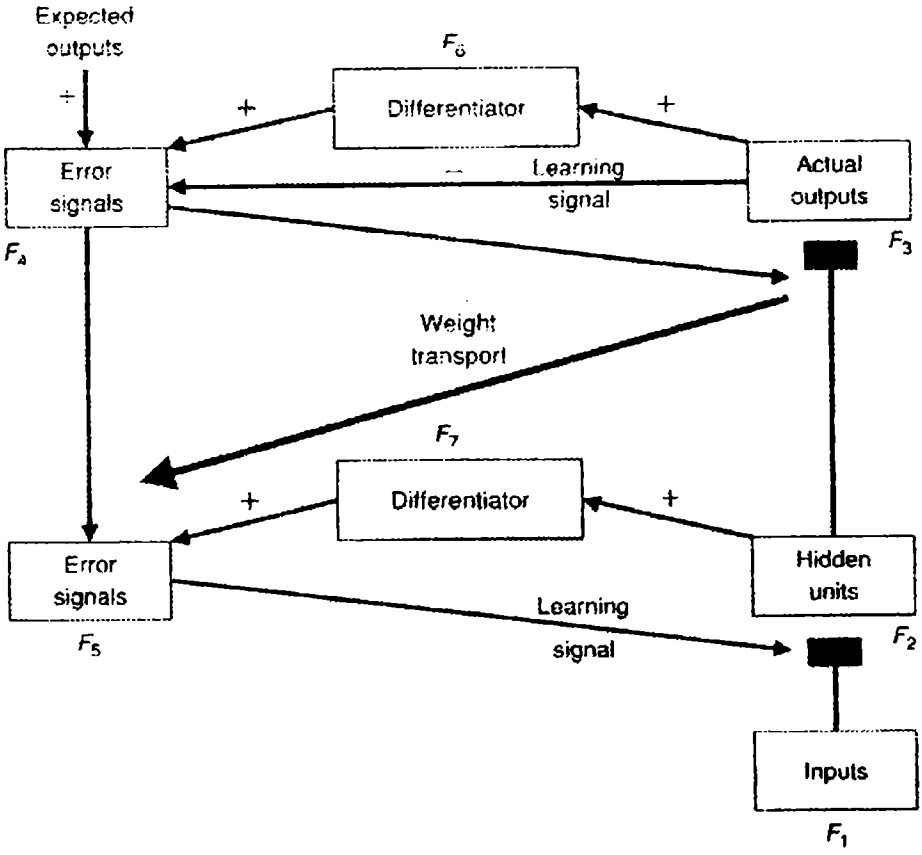


Figura 6.4 *Fluxurile procesării datelor în modelul propagării inverse a erorii [16].*

6.1.6 Aplicații ale rețelelor neuronale

6.1.6.1 Noțiuni fundamentale

Rețelele neuronale constituie, în esență, un mediu specific de procesare a informației în sisteme care operează cu date de intrare mutltivariate şi răspunsuri multiple. *Datele multivariate* sunt asociate cu *un obiect* caracterizat de mai multe variabile; ca exemplu, analiza unor probe de apă potabilă pentru determinarea poluanților Pb, Cl_2 şi CO constituie o problemă cu date de intrare multivariate. *Răspunsurile multiple* sunt asociate cu date multivariate studiate în funcție de mai mulți

https://biblioteca-digitala.ro / https://unibuc.ro

factori care pot influența răspunsul sistemului; în exemplul de mai sus factorii ar putea fi temperatura, presiunea, etc.

Relația generală dintre intrarea m-variat $X_s = (x_{s1}, x_{s2}, ..., x_{sm})$ și răspunsul n-variat $Y_s = (y_{s1}, y_{s2}, ..., y_{sn})$ al sistemului, unde $s = 1, 2, ..., r$ indică eșantioane (probe) diferite, poate fi explicitat compact sub forma:

$$\mathbf{Y}_s = \mathbf{A}\mathbf{X}_s \tag{6.17}$$

unde **A** este în general un operator complex. In particular, **A** poate fi o matrice $(m \times n)$ care transformă liniar vectorul X_s în vectorul Y_s, astfel încât ecuația (6.13) constituie *modelul matematic al problemei multivariate multirăspuns.*

In principiu, problema este următoarea: se dă un set de date m-variate $\{X_s\}_{s=1,2,...,r}$ care conduce la setul răspunsurilor n-variate $\{Y_s\}_{s=1,2,...,r}$, în timp ce operatorul **A** este de cele mai multe ori necunoscut. Mai mult, în majoritatea aplicațiilor se solicită estimarea chiar a răspunsurilor posibile $\{Y_s\}_{s=1,2,...,r}$. Sarcina fundamentală este de a trage concluzii relevante din probe necunoscute pe baza răspunsurilor obținute pornind de la un set de date multivariate selecționate, $\{Y_s\}_{s=1,2,...,p}$, unde $s \in \wp$, $\text{card}\,\wp = p < r$, $\wp$ fiind denumit *setul de antrenament, învățare* sau *instruire.* Nu există un algoritm universal care să selecționeze din mulțimea perechilor de vectori intrare-ieșire $\{(X_s, Y_s)\}$, numite pe scurt configurații, acele perechi care sunt cele mai relevante prin conținutul informațional în procesul de instruire a rețelei.

Alegerea unei strategii de instruire *supervizat* sau *nesupervizat* a rețelei depinde de specificul problemei și datele disponibile. Deoarece instruirea nesupervizată este mai flexibilă prin multitudinea răspunsurilor posibile, este recomandată ca etapă inițială de acumulare a informației. Ulterior, se trece la instruirea supervizată, proces în care datele multivariate cunoscute se împart în trei clase:

- *setul de instruire* (*învățare*) $\wp$, care conține mulțimea configurațiilor de referință;
- *setul de control*, care determină încheierea procesului de instruire;

https://biblioteca-digitala.ro / https://unibuc.ro

- *setul de testare*, care permite verificarea predictibilității rețelei.

Deşi alegerea aleatorie a configurațiilor pentru cele trei clase considerate nu este cea mai indicată strategie, totuşi este singura opțiune rezonabilă în absența unor date sau informații prealabile.

Importanța remarcabilă a selecției configurațiilor în setul de instruire a condus la apariția unui domeniu de specializare denumit *proiectare experimentală* (*experimental design*) sau *proiectare factorială* (*factorial design*). Este foarte important pentru setul de instruire să acopere integral şi cât mai reprezentativ spațiul variabilelor. Un mod consacrat de abordare a acestei probleme constă în selecția variabilelor relevante şi divizarea domeniilor lor de valori în nivele fixe sau intervale limitate (de exemplu: scurte, medii si lungi). Apoi datele sunt selecționate astfel încât setul de instruire să conțină valori ale tuturor variabilelor reprezentând combinații ale tuturor nivelelor sau intervalelor. Reluând exemplul de mai sus, dacă 3 nivele (intervale) posibile de valori sunt asignate la 4 variabile, atunci datele de instruire trebuiesc selecționate astfel încât fiecare din cele 3^4 subspații să conțină cel puțin datele unui experiment. Când numărul de variabile este prea mare pentru o proiectare experimentală cu trei nivele, numărul datelor se reduce, dar se menține acoperirea integrală a spațiului cât mai reprezentativ posibil. In acest sens, se ține seama că variabilele strâns corelate conțin informație relativ similară şi contribuie numai la creşterea redundanței, astfel încât un mod efectiv de a reduce datele de instruire este de a reduce numărul variabilelor la cele care conțin o cantitate maximă de informație nouă.

Limitările majore ale proiectării experimentale sunt două:

- reducerea numărului variabilelor este laborioasă şi conduce inevitabil la pierderea de informație,

- definirea intervalelor fixe pentru domeniul valorilor variabilelor este în bună masură subiectivă şi arbitrară. Rețelele Kohonen oferă o alternativă optimă, în sensul că selecția setului de instruire se realizează prin antrenamentul rețelei Kohonen utilizându-se toate datele de intrare, în timp ce la proiectarea experimentală numărul variabilelor

https://biblioteca-digitala.ro / https://unibuc.ro

este micşorat pentru a reduce numărul subspaţiilor dependente exponenţial de numărul variabilelor.

In cele mai multe aplicaţii, arhitectura care se dovedeşte optimă corespunde reţelelor multistratificate de tip multiperceptron cu un singur strat intermediar (ascuns), la care numărul unităţilor intermediare se stabileşte experimental. O regulă empirică determinată pe mai multe clase de aplicaţii sugerează că numărul optim de unităţi intermediare este primul întreg care depăşeşte logaritmul în baza doi din numărul unităţilor stratului de intrare al reţelei.

La alegerea strategiei adecvate de instruire, factorul imediat important îl constituie numărul elementelor (cardinalul) celor trei mulţimi considerate. Instruirea supervizată poate implica parcurgerea de câteva mii de ori a configuraţiilor de instruire. Mai mult, pot fi necesare corecţii a mii de coeficienţi sinaptici pentru fiecare obiect de intrare. In consecinţă, eficienţa procesului de instruire depinde de numărul configuraţiilor cunoscute, de numărul variabilelor asociate fiecărei configuraţii care defineşte dimensiunea spaţiului (stratului) de intrare, şi de numărul straturilor intermediare (ascunse), care permit reprezentări interne ale informaţiei crescând semnificativ puterea de separare a reţelei.

În caracterizarea performanţelor reţelelor neuronale doi factori sunt consideraţi esenţiali şi anume :

- *reproducerea din memorie* (*"recall"*), care se exprimă prin procentajul configuraţiilor recunoscute corect din *setul de instruire*, după ce procesul de învăţare s-a încheiat,

- *predicţia* (*prediction*), care se exprimă prin procentajul configuraţiilor recunoscute corect din *setul de testare*, configuraţii neutilizate în procesul de instruire a reţelei.

6.1.6.2 Clase de aplicaţii specifice ale reţelelor neuronale

Clasificarea datelor multivariate sortează configuraţiile în clase simple, sau ierarhizează în interiorul claselor. *Ierarhizarea* implică o secvenţă a nivelelor

https://biblioteca-digitala.ro / https://unibuc.ro

decizionale care să clasifice adecvat o configurație într-un grup sau subgrup, categorie sau subcategorie, clasă sau subclasă.

Specific aplicațiilor în chimie, un obiect poate aparține simultan mai multor clase distincte, astfel încât rezultatul este produsul logic al mai multor decizii (fig. 6.5).

Modelarea urmărește elaborarea de sisteme formale capabile să genereze răspunsuri continue pentru datele de intrare, în sensul că la modificări mici la intrare corespund modificări mici ale rspunsului sistemului. Prin esența sa, modelarea necesită întotdeauna instruire supervizată și, în consecință, implică algoritmi de propagare invers sau contra-propagare a erorii (*backpropagation* și *counterpropagation*).

In general, criteriul alegerii parametrilor de modelare este minimizarea diferenței dintre setul răspunsurilor determinate experimental, $\left\{Y_s^{\text{experimental}}\right\}_{s\in\wp}$, și răspunsurile generate de model, $\left\{Y_s^{\text{model}}\right\}_{s\in\wp}$, corespunzătoare aceluiași set de configurații de intrare $\left\{X_s\right\}_{s\in\wp}$, adică:

$$\boldsymbol{min}\left(Y_s^{\text{experimental}} - Y_s^{\text{model}}\right)^2 \qquad (6.18)$$

Avantajul major al rețelelor neuronale în problemele multirăspuns asupra modelării clasice cu funcții analitice constă în faptul că "țintele", $\left\{Y_s^{\text{model}}\right\}_{s\in\wp}$, pot fi multidimensionale. Cu alte cuvinte, modelele clasice necesită predefinirea unei funcții pentru fiecare rspuns, în timp ce rețelele neuronale pot modela un vector multirăspuns fără informații apriorice.

https://biblioteca-digitala.ro / https://unibuc.ro

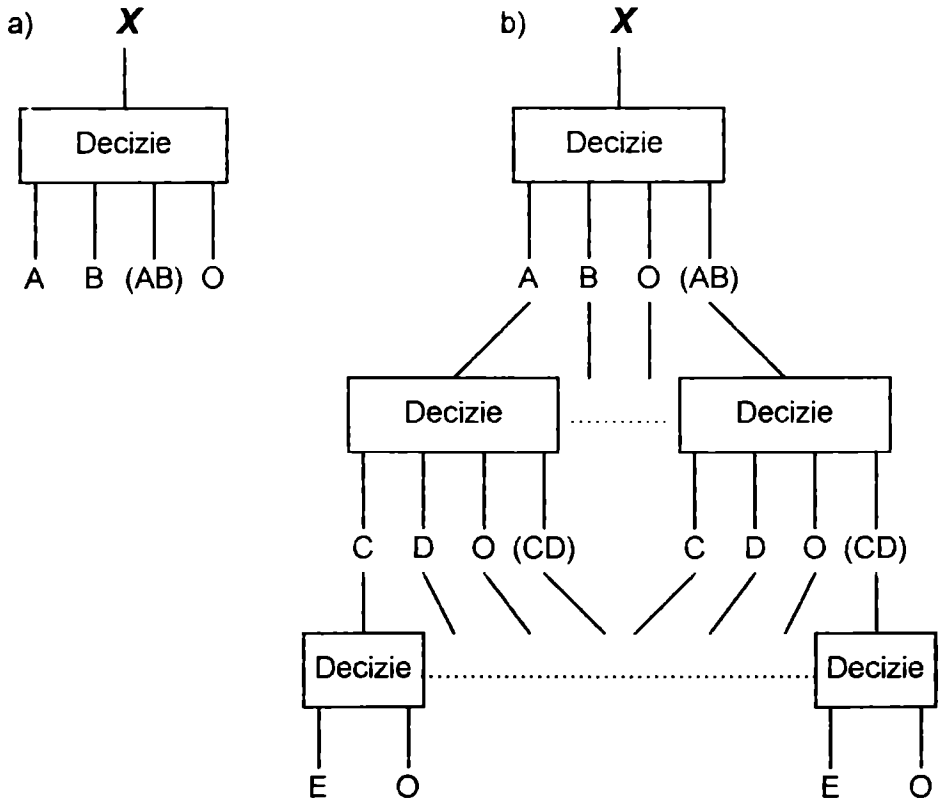


Figura 6.5 *Clasificare pe un nivel decizional (a) şi clasificare ierarhic (b)*

Modelarea este eficientă pentru configuraţii cu un număr relativ redus de variabile de intrare (în general < 10), care conduc la arhitecturi de reţele relativ limitate şi putere de calcul solicitat inferior celor implicate de tehnicile clasice de clasificare (fig. 6.6).

Alegerea algoritmului de instruire este consecinţa direct a distribuţiei obiectelor de intrare în spaţiul variabilelor. Metoda propagării inverse a erorii (*backpropagation algorithm*) este indicat în cazul aplicaţiilor cu un număr relativ redus de obiecte şi repartizate relativ egal. Introdus de Werbos în 1974 şi dezvoltat de Rumelhart şi McClelland [17] este cel mai cunoscut şi utilizat algoritm de învăţare supervizată. El se bazează pe minimizarea diferenţei dintre ieşirea dorită şi ieşirea reală, prin metoda gradientului descendent. Metoda contra-propagării erorii (*counterpropagation algorithm*) este adecvat în cazul unui număr mare de obiecte de intrare repartizate inegal în spaţiul variabilelor, după ce au fost în prealabil reduse la o acoperire regulată a

https://biblioteca-digitala.ro / https://unibuc.ro

spațiului prin intermediul rețelelor Kohonen sau evaluări statistice ale intervalelor variabilelor.

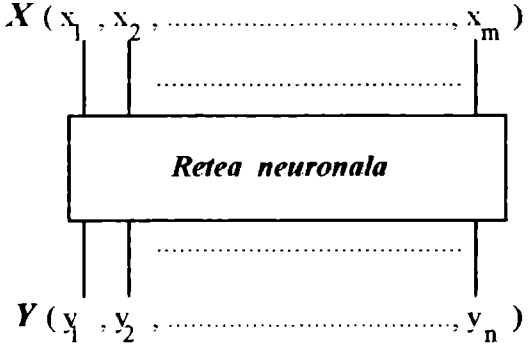


Figura 6.6 *Modelarea cu rețele neuronale*

Mapping*-ul* este în esență o metodă de transformare a unui spațiu de dimensiune mare într-unul cu dimensiune relativ mai mică. In timp ce clasificarea sau determinarea de clusteri se concentrează asupra clasei sau clusterului căruia aparține o anumită configurație, mapping-ul priveste întreg spațiul și efectiv obține clusteri din datele de intrare. Specificitatea rețelelor neuronale constă în reprezentarea celui mai activ neuron prin coordonatele sale și nu prin valoarea ieșirii sale. Elementele esențiale în determinarea corespondenței se referă la topologia spațiului întreg, numărul regiunilor care pot fi distinse, forma lor și corelația dintre caracteristicile spațiului și cele ale configurațiilor de intrare care le-au generat.

Mapping-ul este în particular recomandat ca o metodă de preprocesare și selecție a celei mai adecvate submulțimi de configurații ale unei mulțimi mari, cum ar fi selecția spectrelor pentru corelații spectru-spectru, sau eliminarea combinațiilor redundante de parametri aparținând unei serii de spectre. Retelele Kohonen rezolvă orice problemă posibilă de mapping.

Auto- și ***hetero-asocierea***, deși apropiate ca tehnici de clasificare, pot fi considerate distinct ca metode specifice de identificare corectă a perechilor intrare-ieșire. Rețelele *auto-asociative* furnizează răspunsuri corecte la date de intrare incomplete, distorsionate sau zgomotoase, asociind datelor de intrare $\{X_s\}_{s=1,2,...,m}$

https://biblioteca-digitala.ro / https://unibuc.ro

răspunsuri adecvate $\{Y_s\}_{s=1,2,\ldots,n}$. Rețelele *hetero-asociative* realizează corespondența "unu-la unu" dintre mulțimea obiectelor de intrare $\{X_s\}_{s=1,2,\ldots,m}$ și răpunsurile corespunzătoare $\{Y_s\}=\{X_s\}$, unde s = 1,2,...,m.

Memoriile asociative se instruiesc prin exemple, astfel încât numai sistemele neuronale formale cu instruire supervizată sunt adecvate pentru implementarea funcțiilor asociative.

Prognoza secvențelor temporale modelează comportamentul sistemelor multirăspuns corespunzător unor serii de date înregistrate în timp. Variabilele de intrare și ieșire sunt esențial similare în modelarea temporală; deosebirea este că datele de intrare conțin valorile variabilelor de proces din trecut și prezent, în timp ce datele de ieșire prognozează valorile viitoare ale aceloraşi variabile. Seturile consecutive de date privite ca vectori de intrare și seturile consecutive de date viitoare privite ca vector de ieșire sunt denumite *orizonturi trecute* și respectiv *orizont prezent* ale procesului de instruire [18]. Deși orizonturile pot principial avea orice lungime, totuși orizontul viitor acoperă numai un pas înainte în majoritatea aplicațiilor de control al proceselor.

Fie ***P*** un vector *m*-dimensional care conține informație asupra unui proces dat la un moment de timp *t*:

$$\mathbf{P}_t = (x_{1t}, x_{2t}, \ldots, x_{mt}) \tag{6.19}$$

Dacă se consideră trei pași consecutivi ai procesului, atunci vectorul de intrare este:

$$\mathbf{X} = (x_{1t}, x_{2t}, \ldots, x_{mt}, x_{1t+1}, x_{2t+1}, \ldots, x_{mt+1}, x_{1t+2}, x_{2t+2}, \ldots, x_{mt+2}) \tag{6.20}$$

care se mai poate scrie compact, în termeni vectoriali temporali:

$$\mathbf{X} = (\mathbf{P}_t, \mathbf{P}_{t+1}, \mathbf{P}_{t+2}) \tag{6.21}$$

Vectorul țintă corespunzător ***Y*** poate fi constituit de al patrulea vector în secvență, și anume:

$$\mathbf{Y} = (x_{1t+3}, x_{2t+3}, \ldots, x_{mt+3}) \tag{6.22}$$

sau vectorial:

https://biblioteca-digitala.ro / https://unibuc.ro

$$Y = P_{t+3} \tag{6.23}$$

Deşi modelul complet al unui proces necesită ca toate variabilele de proces să fie rulate, numai un număr limitat de variabile semnificative sunt utilizate la instruire pentru un proces dat. Datele "viitoare" solicitate la iniţierea instruirii sunt obţinute fie din modele teoretice sau realmente din experienţe anterioare.

Investigarea proceselor temporale poate fi utilizată în diverse probleme care implică secvenţe temporale, începând cu prognoza parametrilor de mediu în uzinele chimice şi până la deducerea structurii secundare a proteinelor bazată pe secvenţa aminoacizilor.

6.1.6.3 Abordarea modelării neuronale

Utilizarea reţelelor neuronale este deschis unei largi varietăţi de aplicaţii. In general, tipuri diferite de probleme necesită diferite arhitecturi neuronale şi strategii de instruire adaptate [16].

Ca modalitate principială de abordare a unei probleme utilizând reţele neuronale, este în general indicat a se clasifica mai întâi problema de interes într-una din clasele de aplicaţii prezentate mai sus. In al doilea rând se alege şi se adaptează topologia reţelei care corespunde aplicaţiei identificate. In al treilea rând se optează pentru un algoritm de instruire adecvat capabil să rezolve reţeaua. Indicaţii generale privind clasificarea tipurilor de probleme care se pot aborda cu tehnici de inspiraţie neuronal sunt prezentate pe larg în literatura de specialitate [19-21]. Problemele de chimie conţin adesea elemente care aparţin la tipuri diferite de probleme, astfel încât rezolvarea globală se desfşoară utilizând mai multe tipuri de reţele simultan. Deoarece reţelele neuronale sunt metode generale de rezolvare, un accent deosebit trebuie pus pe reprezentarea informaţiei chimice în context. Cu toate acestea, cele mai multe aplicaţii practice în chimie publicate în literatura de specialitate [22, 23] dovedesc succesul reţelei directe, complet conectate, de tip perceptron, cu un singur strat intermediar (fig. 6.7) şi care se instruieşte utilizând algoritmul de propagare inversă a erorii.

Atunci când reţelele directe clasice sunt instruite pentru mai multe configuraţii cu reguli de propagare derivate din algoritmul de propagare inversă a erorii, pot apărea

https://biblioteca-digitala.ro / https://unibuc.ro

interferenţe puternice la stabilirea ponderilor sinaptice. Aceste interferenţe au ca efect descreşterea convergenţei procesului de instruire şi scăderea puterii de generalizare a reţelei. O cale de a evita interferenţele pentru anumite subclase de configuraţii constă în utilizarea simultană a mai multor *reţele locale diferite* şi o *reţea-poartă* care să decidă reţeaua local activ pentru fiecare subclasă în parte [24]. Ultimile realizări în această direcţie se referă la elaborarea unor noi *funcţii obiectiv* specializate pentru asociaţii adaptive de reţele neuronale locale. Mai clar, în locul unei combinaţii liniare a ieşirilor furnizate separat de reţelele locale, este sugerată abordarea unei reţele-poartă care decide statistic reţeaua local responsabilă pentru clasificarea fiecărcia dintre configuraţiile de intrare [25].

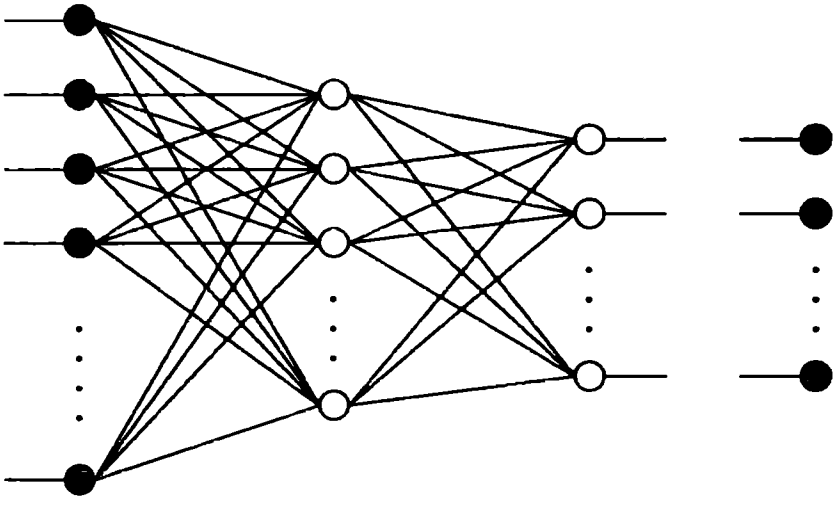


Figura 6.7 *Reţea neuronală de tip perceptron cu un singur strat intermediar şi unităţi ţintă care permit instruirea supervizată*

In chimia analitică mai mult de 90 % din aplicaţii se bazează pe utilizarea algoritmului de instruire supervizată bazat pe propagarea inversă a erorii.

https://biblioteca-digitala.ro / https://unibuc.ro

6.1.6.4 Aplicații ale rețelelor neuronale in chimie

Abilitatea de a procesa date indiferent de caracterul lor (discrete, continui sau vagi), fără specificarea unor corelații prealabile, recomandă rețelele neuronale în aplicații complexe, indiferent de natura datelor. Mai mult, acolo unde nu există modele analitice liniare, rețelele neuronale se dovedesc eficiente datorită caracterului neliniar al răspunsurilor furnizate.

Aplicațiile rețelelor neuronale în științele naturale, și în mod special în chimie, oferă noi metode de cercetare care pot îmbunătăți sau releva noi perspective în modelarea, prelucrarea și interpretarea datelor experimentale. Creșterea continuă a publicațiilor anuale dedicate cercetărilor în domeniul rețelelelor neuronale, ca și a diverselor aplicații raportate, dovedește interesul crescând al comunității științifice pentru sistemele artificiale de inspirație neuronală.

Reactivitatea legăturilor chimice

Reacțiile chimice sunt inițiate prin desfacerea unor legături dintre molecule de aceea cunoașterea reactivității legăturilor chimice este o problemă fundamentală în chimie.

Parametrii care influențează desfacerea legăturilor chimice sunt limitați în modelare la următorul set reprezentativ:

- Δq_{tot} : diferența de sarcină totală,
- Δq_{π} : diferența în sarcină π,
- $\Delta \chi_{\sigma}$: diferența în electronegativitate σ,
- Q_{σ} : măsura polarității legăturii,
- α_b : polarizabilitatea legăturii,
- *BDE* : energia de disociere a legăturii,
- R^{+} : măsura stabilității rezonanței disponibil pentru particulele generate prin heteroliză.

https://biblioteca-digitala.ro / https://unibuc.ro

Valorile acestor parametrii sunt calculate utilizând un pachet de programe numit *PETRA* (*P*arameter *E*stimation for the *T*reatment of *R*eactivity *A*pplications) [21]. Modelul trebuie să coreleze valorile parametrilor cu reactivitatea legăturilor de interes. O abordare eficientă a acestei probleme constă în combinația unei rețele Kohonen pentru selectarea setului de instruire, cu o rețea multistratificată echipată cu instruire supervizată și antrenată pentru clasificare. Ansamblul celor două rețele este capabil sa prezică în moleculele alifatice care dintre legăturile simple se vor desface preferențial polar și care nu, ba mai mult, indică și direcția în care se vor deplasa sarcinile dupa heteroliză. Deși rețeaua poate fi antrenată numai pe legături simple, ea dispune de caracteristica de generalizare, în sensul că asociază reactivitatea corectă inclusiv la unele legături duble (de exemplu C=O). Răspunsurile rețelei pot fi interpretate ca *probabilități* de desfacere a legăturilor, ca și o *măsuăr cantitatiăv* a reactivității.

Optimizarea proceselor HPLC

Metodele cromatografice joacă un rol important în *tehnicile de separare.* Cromatografia de lichide de înaltă performanță (HPLC) constituie o metodă răspândită de analiză în care un amestec este distribuit între o fază staționară, lichidă sau solidă, pe un suport solid, și o fază lichidă mobilă supusă la presiune ridicată. Diversele componente se distribuie diferit între cele două faze și astfel sunt separate. Principalul factor în acest caz îl constituie *factorul de performanță* (*FP*) al analizei, care include două aspecte: puterea de separare a coloanei și timpul necesar pentru desfășurarea integrală a procesului de separare.

Orice analiză de proces HPLC implică modelarea fazei mobile, în sensul specificării factorului de performanță ca funcție de parametrii săi. Apoi se alege cea mai potrivită fază mobilă pentru o analiză particulară.

In majoritatea aplicațiilor se consideră numai doi parametri de care depinde faza mobilă: *concentrația de etanol* (C_e) *și aciditatea fazei* (*pH*). Orice modelare, clasică sau cu rețele neuronale, necesită date experimentale. In modelarea clasică, pentru calculul valorii factorului de performanță se utilizează în general o formă biliniară:

https://biblioteca-digitala.ro / https://unibuc.ro

$$FP = ax^2 + by^2 + cxy + dx + ey = f \tag{6.24}$$

unde x semnifică valorile concentrației C_e, iar y valorile pH-ului. Parametrii a, b, c, d și f sunt obținuți prin orice tehnică de optimizare, cum ar fi *regresia liniară*, metoda *Simplex*, etc.

Aceleași date pot fi utilizate pentru proiectarea unei rețele neuronale formale cu un singur strat intermediar, echipat cu o regulă de instruire de tipul propagării inverse a erorii. Alegerea acestui algoritm de instruire este consecința arhitecturii multistratificate a rețelei, precum și consecința necesității instruirii supervizate reclamată de modelare. Corespunzător proprietăților avute în vedere la modelare, rețeaua trebuie să conțină doi neuroni de intrare, un neuron de ieșire și un număr de neuroni intermediari.

Experiențe de optimizare a proceselor HPLC în analiza vinului au relevat că cele mai performante rețelele de tipul (2×6×1) conținând (2×1)×6+(6+1)×1=25 coeficienți sinaptici. Rezultatele obținute s-au dovedit în bun acord cu cele furnizate prin metode clasice.

Relația structură-activitate

Unul dintre domeniile cele mai promițătoare de aplicații ale tehnicilor de modelare îl constituie determinarea relațiilor cantitative dintre structura chimică a unui compus și activitatea sa fizică, chimică sau biologică. In mod deosebit, corelarea structurii chimice a medicamentelor cu efectele farmacologice este de mare interes în industria farmaceutică. Datorită complexității care poate fi atinsă în modelarea rețelelor neuronale, acestea joacă un rol din ce în ce mai important în studiul relații cantitative structură-activitate, problemă denumită în literatura de specialitate QSAR (*"Quantitative Structure-Activity Relationships"*).

Un exemplu remarcabil în acest sens îl constituie lucrările lui Aoyama, Suzuky și Ichikawa [20] care urmăresc prognozarea dozei medicamentoase minime necesară pentru prelungirea cu 40% a vieții cobailor test, cărora li s-au inoculat celule cancerigene L-1210. *Doza efectivă minimă* depinde de *concentrația C* a substanței care

https://biblioteca-digitala.ro / https://unibuc.ro

produce efectul dorit şi se exprimă în termeni logaritmici ($\log(1/C)$), consecinţă a plajei de mai multe ordine de mărime a concentraţiilor solicitate.

Setul de date cuprinde modificările caracteristicilor structurale ale carbochinonelor, o clasă de compuşi care manifestă activitate *anticancerigenă* cu diverse intensităţi. Această activitate este în primul rând influenţată de identitatea substituenţilor R^1 şi R^2. In analiza multiliniară standard, influenţa combinată a substituenţilor R^1 şi R^2 este descrisă de următoarele variabile fizico-chimice:

- $MR_{1,2}$: refractivitatea molară,
- $\pi_{1,2}$: contribuţia substituenţilor la hidrofobicitate,
- F : suma constantelor substituenţilor la efectul de câmp,
- R : suma constantelor substituenţilor la efectul de rezonanţă; şi două variabile locale responsabile de influenţa exclusivă a lui R^1:
- MR_1 : refractivitatea molară a lui R^1,
- π_1 : contribuţia lui R^1 la hidrofobicitate.

Au fost selecţionate 35 de carbochinone diferite pentru care, fie au fost măsurate valorile celor 6 variabile ($MR_{1,2}$, $\pi_{1,2}$, MR_1, π_1, F şi R), fie au fost preluate date din literatura de specialitate (π_1, F şi R), apoi doza efectiv minimă a fost măsurată şi reprezentată ca $\log(1/C)$. In consecinţă, a fost utilizată o reţea neuronală cu 6 unităţi de intrare corespunzătoare celor 6 variabile ($MR_{1,2}$, $\pi_{1,2}$, MR_1, π_1, F şi R), o unitate de ieşire şi o unitate ţintă, care specifică doza minimă efectivă $\log(1/C)$. In majoritatea aplicaţiilor s-a dovedit suficient un singur strat intermediar al cărui număr de unităţi s-a stabilit experimental. In aceste condiţii, a fost utilizată o variantă a regulii de instruire cu propagare inversă a erorii. Rezultatele obţinute cu reţele neuronale sunt superioare, deşi nu în mod dramatic, celor furnizate de tehnica clasică MLRA (*"Multi-Linear Regression Analysis"*).

https://biblioteca-digitala.ro / https://unibuc.ro

Caracteristicile generale ale structurii secundare a proteinelor

Determinarea structurii macromoleculelor biologice este o problemă fundamentală în biologia moleculară şi constituie un prim pas în elucidarea proprietăţilor funcţionale complexe care le caracterizează. Forma elementară a problemei priveşte prognoza structurii secundare a proteinelor, adică conformaţia spaţială la nivel inferior (de exemplu α-*elicii*, β-*plane* şi *spirale aleatorii*) pe baza structurii lor primare [26]. In această problemă, legea care determină structura secundară a lanţului proteic în funcţie de propritătile colective ale aminoacizilor individuali este principala necunoscută. Modurile de abordare a problemei se încadrează principial în două categorii: *tehnici clasice* şi *metode specifice inteligenţei artificiale*.

Abordarea neuronală a modelării structurii secundare din perspectiva tehnicilor aparţinând domeniului inteligenţei artificiale este superioară deoarece reţelele neuronale sunt caracterizate de o procesare holistică a datelor de intrare şi, mai mult, nu necesită reguli de instruire *ad hoc* sau informaţii extrem de *specifice*, ca o consecinţă directă a procesului de instruire prin exemple şi a puterii de generalizare a reţelei.

Reţelele neuronale recomandate în acest gen de aplicaţii sunt de tip perceptron multistratificat, iar algoritmul de instruire de tipul cu propagare inversă a erorii. Configuraţia unei secvenţe primare de reziduuri de lungime l este aplicat la intrarea perceptronului, care urmează să asocieze aminoacidului din mijlocul configuraţiei (l este presupusă impară) structura sa secundară. Dimensiunea l a ferestrei este astfel aleasă încât reziduurile configuraţiilor de intrare să cuprindă cele mai multe din interacţiile care prezumptiv ar stabiliza structura. Experimental, s-a determinat că dimensiuni ale ferestrei mai mari de 5 reziduuri nu aduc în mod necesar o îmbunătăţire sensibilă, dar s-a preferat în aplicaţii valoarea 7. Studiul statistic al proteinelor a relevat că lungimea unei configuraţii afectează moderat performanţa metodei. A fost utilizat un cod expandat, în care grupuri de câte 20 neuroni au fost folosiţi la codificarea fiecărui aminoacid din cei 20. Dintre aceştia, 19 aminoacizi sunt inactivi, iar poziţia celui de-al 20-lea neuron activ în interiorul grupului identifică reziduurile considerate. Astfel se

https://biblioteca-digitala.ro / https://unibuc.ro

evită corelațiile nedorite între diverse reziduuri care sunt simultan prezente la intrarea perceptronului. In consecință, numărul neuronilor de intrare este $7 \cdot l$. Fiecare neuron de ieşire reprezintă una din posibilele structuri a fi prognozate şi ieşirea rețelei este făcută să corespundă cu neuronul având activitate maximă. La instruire au fost necesare cel puțin 100 de cicluri complete de parcurgere a setului de configurații de referință [25].

În concluzie rețelele neuronale testate pe clase *omogene* de proteine(setul de instruire şi cel de testare sunt alese din aceeaşi clasă), fie globulare, fie membranare, manifestă o putere de generalizare superioară metodelor clasice de prognoză. Mai mult, rețelele de tip perceptron sunt capabile de a identifica şi alte structuri secundare decât elicii tramsmembranare (de exemplu elicii extramembranare sau β-spirale). În al doilea rând, experiențele efectuate pe clase de proteine presupuse *eterogene* (setul de instruire şi cel de testare sunt alese din clase diferite) au evidențiat abilitatea perceptronilor de a învăța şi recunoaşte configurații şi proprietăți statistice cu o eficiență comparabilă cu cea obținută pe clase omogene. În al treilea rând, superioritatea sensibilă a performanței sistemelor neuronale de tip perceptron în raport cu metodele clasice aplicate proteinelor membranare, sugerează că metodele de abordare cu rețele neuronale sunt mai eficiente în identificarea regularităților statistice conținute în setul de instruire, precum şi mai eficiente decât metodele statistice în extrapolarea datelor structurale de la proteinele globulare la proteinele membranare [3, 27].

Corelația spectru-structură

Elucidarea structurii compuşilor organici se bazează masiv pe metode spectroscopice, deşi relațiile dintre structură şi datele spectrale sunt în general prea complexe pentru a fi exprimate sub forma unor ecuații explicite. Se cunosc numeroase tehnici de abordare a interpretării automate a spectrelor infraroşii printre care: căutarea şi fitarea cu spectrele aparținând unei baze de date, utilizarea tabelelor de corelații, diferite metode statistice, teoria mulțimilor, precum şi diverse metode de recunoştere de forme (*"pattern recognition"*) şi sisteme expert (*"expert systems"*) [28]. Corelațiile precise dintre spectrele complete de înaltă rezoluție şi structurile chimice asociate,

https://biblioteca-digitala.ro / https://unibuc.ro

precum şi disponibilitatea unor largi baze de date computerizate, fac extrem de atractiv acest domeniu pentru aplicaţii de interes a reţelelor neuronale.

Sistemele neuronale formale constituie o clasă de metode elegante de clasificare a configuraţiilor în care un vector de intrare, ale cărui valori reprezintă obiectul de intrare, este transformat într-un vector de ieşire, ale cărui valori reprezintă clasificarea dorită a obiectului de intrare. Transformarea este realizată prin intermediul unei formule de propagare, care combină vectorul de intrare cu coeficienţii interni ai reţelei. Obiectele de intrare, care constituie subiectul transformării, sunt spectrele infraroşii ale unor compuşi diverşi, în timp ce ieşirea furnizează serii de substructuri care sunt conţinute (sau nu) de compusul reprezentat la intrare. Procesul include *interpretarea spectrală*, adică exprimarea proprietăţilor spectroscopice ca fragmente substructurale, şi *generarea structurii*, adică reuniunea fragmentelor substructurale alcătuind toate structurile moleculare complete compatibile cu ele.

O tehnică convenabilă de a opera cu date complexe multivariate este *clasificarea*, care sortează obiecte în clase de echivalenţă sau ierarhizează subclase în clase. Uzual, stabilirea unei clasificări ierarhice în decizii complexe începe cu un studiu statistic prealabil al relaţiilor dintre reprezentarea obiectelor şi clasele lor de apartenenţă. Clasificările ierarhice implică mai multe nivele decizionale şi, mai mult, sunt aplicaţii în care obiectul de intrare este caracterizat imprecis sau vag (*"fuzzy"*). Modelele clasice necesită funcţii analitice predefinite pentru fiecare răspuns, în timp ce reţelele neuronale prezintă avantajul că pot modela vectori multirăspuns fără cunoştinţe apriorice şi procesa atât funcţii de apartenenţă boolene (*"crisp membership functions"*), cât şi funcţii de apartenenţă vagi (*"fuzzy membership functions"*). Analiza corelaţiei spectru-structură constituie o problemă tipică de clasificare multidecizională: obiectul de intrare este configuraţia unui spectru, iar răspunsul este o listă de clase prezente (sau absente) în compusul care a produs spectrul de la intrare. Atât instruirea supervizată, cât şi cea nesupervizată, pot fi utilizate în proiectarea reţelelor neuronale destinate clasificării; în primul caz se recomandă o arhitectură cu un singur strat intermediar

https://biblioteca-digitala.ro / https://unibuc.ro

instruită cu regula de propagare inversă a erorii, în cel de-al doilea caz se face invariabil apel la rețele de tip Kohonen [29].

Munk, Madison şi Rob [30] sugerează diviziunea spectrului infraroşu de la 3960 la 400 cm^{-1} în 640 intervale de 5,6 cm^{-1} lățime. Intensitatea x_i a radiației transmise de intervalul *i* este scalată la:

$$x_i = 1.00 - (\%T)/100.0 \tag{6.25}$$

unde *T* este factorul de transmisie. Numărul mare de unități de intrare a fost redus experimental la 256, pentru a elimina rspunsurile neconcludente. O altă îmbunătățire a metodei constă în îngustarea lățimii intervalelor la joasă frecvență şi lărgirea lor la înaltă frecvență, cu scopul de a le adapta la puterea de discriminare variabilă de la un capăt al altuia al spectrului. In consecință, dependența de frecvență a lungimii intervalului *i* este dată de următoarea expresie:

$$i = E\left(6.0\sqrt{\nu} - 120.0\right) \tag{6.26}$$

unde $E(\cdot)$ este funcția parte întreagă, iar ν este frecvența. Domeniul de frecvență asignat intrării *i* este apoi explorat şi sunt înregistrate vârfurile în amplitudine. Dacă un maxim este prezent într-un interval, atunci amplitudinea sa este scalată între 0.000 şi 1.000, în caz contrar valoarea intrării este setată la *zero*.

Experimental, compuşii au fost descrişi în termenii a 36 grupări funcționale reprezentative (alcooli primari, fenoli, amine, aminoacizi, esteri, etc.), care au fost alese din 140 grupări funcționale studiate anterior [30]. Fiecărei grupări i s-a repartizat o unitate de ieşire arbitrar, astfel încât răspunsul țintă devine un vector 36-variat cu componente binare în care valoarea *1* indică prezența grupării funcționale asociate iar *0* indică absența sa. Pe scurt, esența acestei probleme 256-variate cu 36-răspunsuri este de a stabili prezența sau absența a oricărei grupări funcționale din cele 36 considerate relevante pe baza unor spectre diferite alcătuite din 256 puncte. Majoritatea compuşilor analizați conțin mai multe grupări funcționale, astfel încât multe unități de ieşire sunt simultan active. Prin încercări s-a ajuns la arhitectura optimă a rețelei: (256×34×36). Biblioteca de spectre a cuprins 2915 compuşi, dintre care au fost aleşi aleatoriu 2499

https://biblioteca-digitala.ro / https://unibuc.ro

pentru setul de instruire şi restul de 416 pentru setul de test. A fost respectată astfel o regulă empirică generală referitoare la instruirea reţelelor mari utilizând algoritmul cu propagarea inversă a erorii, şi anume ca numărul datelor alese pentru instruire să depăşească numărul ponderilor sinaptice modificabile ale reţelei. In cazul studiat, setul de instruire conţine aproximativ 650.000 valori (2.499 spectre × 256 intensităţi), în timp ce numărul coeficienţilor modificabili este puţin sub 350.000 (256×34×36).

Reţelele neuronale s-au dovedit, deşi pe un număr restrâns de experimente publicate până în prezent, a fi extrem de promiţătoare în competiţia cu alte sisteme expert dedicate elucidării structurale a compuşilor chimici prin prelucrare automată a datelor [31]. Mai mult, reţelele neuronale pot fi combinate cu sisteme expert şi implementa o combinaţie de metode care, adesea, se dovedeşte mai tare decât fiecare în parte. Domeniul de explorare a sistemului expert se limitează la compuşi pentru care răspunsurile reţelei nu sunt concludente [32, 33].

Sisteme automate de control al proceselor şi detecţie a defectelor

Două clase de probleme caracterizează desfăşurarea proceselor chimice. Prima clasă priveşte recunoşterea şi detecţia defectelor de o manieră calitativă, atunci când este necesar să se identifice numai disfuncţionalităţile sistemului şi/sau părţile sale defecte. A doua clasă constă în controlul efectiv al procesului însuşi, din moment ce valorile variabilelor de proces trebuiesc continuu monitorizate şi menţinute în limitele prescrise de tehnologie aplicând reacţie negativă (*"feed-back techniques"*).

Controlul fiabil al unui proces dat se realizează dacă sunt satisfăcute două cerinţe fundamentale. Prima priveşte concepţia unui model **M** al procesului în măsură să prezică parametrii critici ai procesului pe durata unei secvenţe de câteva unităţi de timp înainte; a doua cerinţă impune inversabilitatea modelului **M**, adică determinarea modelului $\mathbf{M}^{-1}$, care să permită ajustarea variabilelor de corecţie pe baza prognozei datelor furnizate de modelul direct **M**, astfel încât sistemul (procesul) să revină la normal dacă ar fi aplicate aceste ajustări. Atât algoritmul cu propagare inversă, cât şi

https://biblioteca-digitala.ro / https://unibuc.ro

cel cu contra-propagare a erorii se dovedesc comparabile ca rezultate. Totuşi, algoritmul cu contra-propagare a erorii poate fi privit ca un procedeu de netezire neliniară (*"nonlinear smoothing"*), astfel încât solicită o cantitate de date de intrare mai mare pentru o acurateţe dată, dar timpul de instruire total nu creşte rapid cu dimensiunea setului de intrare. Deoarece numărul ciclurilor complete de antrenament este similar cu cel solicitat de instruirea reţelelor Kohonen, convergenţa în reţelele cu contra-propagare se atinge cu ordine de mărime mai rapid decât în reţelele cu propagare inversă a erorii.

Implementarea sistemelor de control al proceselor şi detecţia defectelor se bazează masiv pe modelarea matematică a proceselor, fapt care atrage dificultăţi suplimentare:

-lipsa unor modele sigure pentru un proces dat adecvat tuturor circumstanţelor care pot apare;

-relaţiile neliniare complexe dintre variabilele de control şi desfăşurarea procesului face extrem de dificilă, dac nu imposibilă dezvoltarea de modele analitice;

-datele de control reactiv zgomotoase, imprecise sau incomplete, afectează negativ eficienţa şi siguranţa monitorizării datelor;

-lipsa unei relaţii generale de izomorfism între setul de simptome observabile şi diagnoza corectă.

Mai multe metode de abordare au fost propuse în cadrul teoriilor de control automat al proceselor:

-analiza statistică;

-modele liniare care utilizează tehnici clasice,

-modele de recunoaştere de configuraţii (*"pattern recognition"*);

-sisteme expert;

-reţele neuronale echipate cu reguli de instruire cu propagare inversă a erorii.

Hoskins and Himmelblau [34] au tratat exhaustiv diversele aplicaţii ale reţelelor neuronale în ingineria chimică, cum ar fi detecţia defectelor (*"fault detection"*), diagnoza (*"diagnosis"*), controlul proceselor (*"process control"*), proiectarea

https://biblioteca-digitala.ro / https://unibuc.ro

proceselor (*"process design"*) şi simularea proceselor (*"process simulation"*). Deşi este oarecum dificil a separa net aceste tipuri de aplicaţii, totuşi reţelele neuronale prezintă global o serie de avantaje semnificative asupra tehnicilor clasice:

-nu sunt necesare modele analitice deoarece reţelele neuronale învaţă din exemple;

-modelele neliniare sunt agreate datorită răspunsurilor inerent neliniare ale reţelelor neuronale;

-funcţiile de asociere şi clasificare caracteristice reţelelor neuronale le recomandă pentru procesarea datelor vagi (*"fuzzy"*), în mod special a celor aflate în afara regiunilor de instruire;

-date de intrare de orice natură pot fi procesate simulatan (continue, discrete şi vagi);

-reţelele neuronale pot modela funcţii inverse, o caracteristică esenţială în generarea semnalelor de control reactiv.

Un proces chimic este monitorizat prin măsurarea continuă a unui număr de variabile, x_i, în diferite puncte ale procesului şi supervizat de semnale de reacţie care acţionează asupra unor elemente de control. Sarcina *elementelor de control* este de a menţine variabilele măsurate într-o stare dată (*activ/inactiv*), fie la o valoare fix (*"set-point"*), fie în interiorul unui interval definit (*"recommended range"*). Ansamblul tuturor variabilelor x_i măsurate şi al stărilor la un moment dat t, este denumit *vectorul de process* $\mathbf{P}_t$:

$$\mathbf{P}_t = (x_{1t}, x_{2t}, \ldots, x_{it}, \ldots, x_{mt}) \tag{6.27}$$

Vectorii de instruire, X, formează mulţimi de variabile măsurate care sunt cunoscute că asigură funcţionarea normală a procesului, iar vectorii ţintă, Y, specifică stările anormale la care poate conduce intrarea X. In general, cauza unei disfuncţionalităţi este o combinaţie de stări anormale care adesea sunt *binare* sau *discrete*. Modul cel mai elegant de a trata *stările discrete multiple* ale unei variabile

https://biblioteca-digitala.ro / https://unibuc.ro

este de a transforma variabila într-o *reprezentare distribuită*, adică fiecare variabilă multistare este transformată în tot atâtea intrări binare câte stări are variabila.

Modul de variație a variabilei de proces x_i este prognozat dacă X conține cel puțin două valori consecutive în timp (de exemplu x_{it} și x_{it+1}). Așadar *orizontul trecut* al vectorului de instruire X trebuie să aibă o lungime minimă de două evenimente în tehnica ferestrei (*"moving window"*), în timp ce următoarea valoare a variabilei, x_{it+2}, trebuie să fie conținută în *orizontul viitor*, Y (vectorul țintă). Dacă previziunile se fac pentru mai multe unități temporale înainte, Y poate avea o dimensiune mai mare de o unitate de timp. Se recomandă ca orizontul trecut să fie cel puțin de două ori mai mare decât cel viitor.

Toate aplicațiile legate de controlul automat al proceselor necesită instruire supervizată bazată pe regula propagării inverse a erorii. Dimensiunea rețelei este în primul rând impusă de numărul variabilelor de intrare și ieșire. O simplificare poate fi introdusă, mai ales în sistemele de control cantitativ, în sensul că nu este necesar ca toți vectorii P_t din setul de instruire să conțină toate variabilele de proces, ci numai pe acelea prin care se realizează efectiv controlul procesului.

6.1.6.5 *Concluzii finale privind aplicațiile rețelelor neuronale in chimie*

Aplicațiile rețelelor neuronale în științele naturale, și în mod special în chimie, oferă noi metode de cercetare care pot îmbunătați sau releva noi perspective în modelarea, prelucrarea și interpretarea datelor experimentale. Creșterea continuă a publicațiilor anuale dedicate cercetărilor în domeniul rețelelor neuronale, ca și a diverselor aplicații raportate, dovedește interesul crescând al comunității științifice pentru sistemele artificiale de inspirație neuronală.

Particularitățile care fac rețelele neuronale atractive sunt multiple și cu ponderi variabile în aplicații specifice. Se impune însă reprezentarea proprie a informației în interiorul rețelei multistratificate, fără echivalent conceptual direct, dar care crește considerabil eficiența rețelei ca rezolvant extrem de general al problemelor de

https://biblioteca-digitala.ro / https://unibuc.ro

clasificare şi clustering. In aceste condiţii, suntem forţaţi să reconsiderăm reprezentarea şi interpretarea datelor experimentale pentru a extrage maximul de informaţie utilă.

Abilitatea de a procesa date indiferent de caracterul lor (discrete, continui sau vagi), fără specificarea unor corelaţii prealabile, recomandă reţelele neuronale în aplicaţii complexe, indiferent de natura datelor. Mai mult, acolo unde nu există modele analitice liniare, reţelele neuronale se dovedesc eficiente datorită caracterului neliniar al răspunsurilor furnizate.

Abordarea datelor de intrare la nivelul întregii reţele le recomand în aplicaţii cu date multivariate corelate, adică exact unde modelarea clasică este mai greoaie şi mai puţin transparentă.

Una din tendinţele actuale de a rezolva problemele hipercomplexe este splitarea lor în subprobleme mai simple reunind în final soluţiile parţiale. Reţelele neuronale sunt capabile de a realiza clasificarea şi asocierea pe un singur nivel şi mai mult, procesează seturi mari de date eficient prin clasificare ierarhizată, utilizând reţele-poartă. Concret în chimia analitică, reţelele neuronale sunt folosite la identificarii, studiul clustărilor, clasificări şi modelări ale datelor chimice sau estimari de parametrii.

Condiţia primordială a succesului în aplicaţii cu sisteme artificiale neuronale rămâne de departe reprezentarea informaţiei în context. Sensibilitatea reţelei la datele de intrare conţinute în setul de instruire este deosebită şi a condus pe mulţi adepţi ai tehnicilor de rezolvare a problemelor cu reţele neuronale la o "regulă de aur" formulată extrem de simplu şi sugestiv: *"garbage in - garbage out"*!

6.2 Algoritmi genetici

Alături de reţelele neuronale, algoritmii genetici [35] studiaţi şi dezvoltaţi de Holland [36] aparţin metodelor care sunt analoage metodelor biologice. In biologie algoritmul genetic se bazează pe evoluţia componenţilor cum ar fi populaţia fiinţelor vii, competiţia între fiinţele vii şi reproducerea vieţii. Astfel algoritmul genetic are la bază un ansamblu de opţiuni iniţiale cum ar fi populaţia de cromozomi în procesul de

https://biblioteca-digitala.ro / https://unibuc.ro

învățare. Caracteristicile genetice ale algoritmilor de instruire se referă la strategiile de reproducere, selecție și mutații. Pentru a aplica conceptele evoluției genetice se introduc *cromozomii*, care codifică informația. Se introduce de asemenea un criteriu de apreciere a diferiților cromozomi corespunzător performanței în raport cu o problemă dată pentru aplicarea conceptului de selecție. Această metodă poate fi utilizată pentru realizarea optimizării unui set de variabile de o înaltă complexitate în sensul unei selecții. Un exemplu tipic de aplicație îl constituie selecția lungimii de undă în analiza spectrometrică multicomponent. Altă aplicație ar fi cea a optimizării parametrilor ca de exemplu estimarea parametrilor în modelele neliniare [37]. De asemenea algoritmul genetic poate fi folosit ca algoritm de învățare pentru rețelele neuronale înlocuind algoritmul cu propagare inversă a erorii. Alături de aplicațiile numerice, algoritmii genetici pot fi folosiți pentru rezolvarea unor probleme combinatoriale. Aplicațiile algoritmilor genetici se caracterizează prin câteva particularități cum ar fi [37]:

- algoritmii operează în baza variabilelor codificate și nu în baza variabilelor de intrare ca atare;
- căutarea optimului pornește din câteva puncte;
- criteriul pentru optimizare este funcția obiectiv și nu derivatele acesteia.

In concluzie, algoritmii genetici nu necesită o alegere inițială a parametrilor problemei ci un întreg ansamblu de opțiuni inițiale numit populația inițială de cromozomi. Această populație este ulterior procesată prin procedee de inspirație biologică reproducând și modificând simultan compoziția sa genetică.

Algoritmul genetic este de asemenea folosit și în alte aplicații care presupun selecția variabilelor or recunoașterea configurațiilor sau în metodele de calibrare [38-40]. Lavine [41] dezvoltă un algoritm genetic pentru explorarea unor baze mari de date pentru identificarea și optimizarea separării acestora pe clase în funcție de doi sau trei componenți principali. In general algoritmul este direcționat către acele clase de compuși care sunt greu de clasificat. Creșterea puterii computaționale în timp va face ca algoritmul genetic să devină mult mai practic pentru o serie de aplicații. Studiile

https://biblioteca-digitala.ro / https://unibuc.ro

privind utilizarea algoritmului genetic în rezolvarea problemelor chimiştilor sunt în atenţia cercetătorilor.

6.3 Analiza imagistică

Imagistica chimică este o combinaţie între spectroscopia moleculară si imagistica digitală. Seturile de date generate de imagistica chimică sunt multe şi în general multivariate deci este necesară o procesare semnificativă a acestora [42,43]. Schweitzer [44] într-un bine documentat review redă cu eleganţă unele dintre aplicaţiile tehnicilor chemometrice la procesarea şi vizualizarea datelor imagistice. Astfel segmentarea şi clasificarea pot fi utilizate cu succes în procesarea datelor analizelor imagistice. Recente aplicaţii ale chemometriei în analiza imagistică au fost prezentate într-o serie de lucrări prezentate în literatura de specialitate [45, 46]. Una dintre metodele chemometrice considerată eficientă pentru analiza imagistică este algoritmul analizei factoriale propusă pentru studiul imaginilor multivariate [47]. Multe dintre aplicaţiile analizei imagistice continuă să fie direcţionate către imagistica creierului (brain imaging). Buydens (48) studiază acest domeniu cu precădere combinaţia dintre imaginile oferite de către MRI (imagistica de rezonanţă magnetică) şi MRSI (imagistica spectroscopică de rezonanţă magnetică) folosind chemometria pentru obţinerea informaţiilor suplimentare la testul pe subiecţi. Nicholson [49] şi Winding [50] au investigat de asemenea utilizarea recunoaşterii de configuraţii în imagistica rezonanţei magnetice biomedicale. Altă aplicaţie a analizei imagistice include utilizarea spectroscopiei autofluorescente si microscopia electronică pentru studiul ţesutului uman, alimente şi pulberi. In literature de specialitate există o serie întregă de aplicaţii ale analizei imagistice cu precădere în domeniul medical. Acest domeniu de actualitate este în continuă ascendenţă fiind în atenţia cercetătorilor.

Bibliografie

1. E. Cherniak, D. McDermott, *Introduction to Artificial Intelligence*, Addison-Wesley, Reading, MA, 1985.
2. R. Mutihac, L. Mutihac, *Roum. Chem.Quart. Rev.* 3, 329, 1995.

https://biblioteca-digitala.ro / https://unibuc.ro

3. R. Mutihac, L. Mutihac, *Roum. Chem.Quart. Rev.* 1, 67, 1997.

4. P. Smolensky, în *Parallel Distributed Processing, Exploration in the Microstructure of Cognition* (D. Rumelhart şi J McCleland, eds.), The MIT Press,Cabridge, MA, 1986.

5. M. Otto, *Chemometrics, Statistics and Computer Application in Analytical Chemistry,* Wiley-VCH, Weinheim, 1999.

6. J. J. Hopfield, *Proc. Nat. Acad. Sci.* 79, 2554, 1982.

7. R. Mutihac, V. Gheorghe, *Rev. Roum. Phys.* 37, 535, 1992.

8. R. Mutihac, *Romanian Reports in Physics* 47, 76, 1995.

9. L. Mutihac, V. David, *Chemometrie,* Ed. Universităţii din Bucureşti, Bucureşti, 1997.

10. C. Bala, L. Mutihac, *Chemometrie, Principii şi Aplicaţii în Chimia Analitică,* Ed. Universităţii din Bucureşti, Bucureşti, 2000.

11. J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms,* Plenum Press, New-York, 1982.

12. R. Mutihac, *St. Cerc. Fiz.* 43, 603, 1991.

13. R. Mutihac, *Proc. 13th World Congress on Computation and Applied Mathematics,* Dublin, Ireland, p.1009, 1991.

14. F. Rosenblatt, *Principles of Neurodynamics,* Spartan Books, 1962.

15. W. S. McCulloch, W. Pitts, *Bull. Math. Biophys.* 5, 115, 1943.

16. R. Mutihac, *Modelarea şi simularea reţelelor neuronale*, Teză de doctorat, Universitatea din Bucureşti, 1994.

17. D. E. Rumelhart, J. L. McClelland, PDP *Parallel Distributed Processing, Explorations in the Microstructure of Cognition,* The MIT Press, Cabridge, MA, 1986.

18. J. Zupan, *Algorithms for Chemists*, J. Wiley, Chichester, UK, 1989.

19. S. M. Deming, S. L. Morgan, *Experimental Design: A Chemometric Approach,* Elsevier, Amsterdam, NL, 1987.

20. K. Varmuza, *Pattern Recognition in Chemistry*, Springer-Verlag, Berlin, 1980.

21. J. Zupan, J. Gasteiger, *Neural Networks for Chemists*, VCH, Weinheim, 1993.

22. D. L. Massart, B.G. M. Vandeginste, S. N. Deming, Y. Michotte, L.

https://biblioteca-digitala.ro / https://unibuc.ro

Kaufmann, *Chemometrics: A Textbook*, Elsevier, Amsterdam, NL, 1988.

23. V. Kvasnièka, *Neur. Net. World* 2, 171, 1993.

24. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, *Neur. Comp.* 3, 79, 1991.

25. W. Kabsch, C. Sander, *Dictionary of Protein Secondary Structure: Pattern of Hydrogen-Bonded and Geometrical Features*, *Biopolymers* 22, 2577, 1983.

26. M. Compiani, P. Fariselli, şi R. Casadio, in *Third Italian Workshop on Parallel Architectures and Neural Networks* (E.R. Caianiello, ed.), World Scientific Publishing Co., 1990.

27. J. Garnier, D.J. Osguthorpe, B. Robson, *J. Mol. Biol.* 120, 97, 1978.

28. E.W. Robb, M.E. Munk, *Mikrochim Acta* I, 131, 1990.

29. T. Kohonen, *Self-Organisation and Associative Memory*, Springer, Berlin, 1984.

30. M.E. Munk, M.S. Madison, E.W. Robb, *Mikrochim Acta* II, 505, 1991.

31. R. Mutihac, *Modelarea şi Simularea Neuronală*, Ed. Universităţii din Bucureşti, Bucureşti, 2000.

32. W.J. Melssen, J.R.M. Smits, G.H. Rolf, G. Kateman, *Chemom. Intell. Lab. Syst.* 1993.

33. M. Novic, J. Zupan, *Vestn. Slov. Kem. Drus.* 39, 195, 1992.

34. J. C. Hoskins, D. M. Himmelblau, *Comput. Chem. Eng.* 12, 881, 1988.

35. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 1989.

36. J. H. Holland, *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, MI, 1975.

37. M. Otto, *Chemometrics, Statistics and Computer Application in Analytical Chemistry*, Wiley-VCH, Weinheim, 1998.

38. C. B. Lucasius, M. L. M. Beckers, G. Kateman, *Anal. Chim. Acta* 286, 135, 1994.

39. F. K. Hopke, *Anal. Chim. Acta* 500, 365, 2003.

40 B. K. Lavine, D. Brzozowski, A. J. Moores, C . E. Davidson, H. T. Mayfield, *Anal. Chim. Acta* 437, 233, 2001.

41. B. K. Lavine, C. E. Davidson, A. J. Moores, *Chemom. Intell. Lab.* 1, 161, 2002.

42. R. Mutihac, *Introducere în procesarea de imagini*, Ed. Universităţii din Bucureşti, 1999.

https://biblioteca-digitala.ro / https://unibuc.ro

43. R. Mutihac, *Procesarea digitală a imaginilor*, Ed. Universităţii din Bucureşti, 2000.

44. R. C. Schweitzer, A. S. Bangalore, P. J. Treado, *Managing Mod. Lab.* 5, 7, 2000.

45. B. K. Lavine, J. Workman, Jr. *Anal. Chem.* 74, 2763, 2002.

46. P. K. Hopke, *Anal. Chem.* 500, 365, 2003.

47. J.-H. Wang, P. K. Hopke, *Chemom.Intell. Lab. Syst.* 55, 13, 2001.

48. H. Witjes, A. W. Simonetti, L. Buydens, *Anal. Chem.* 73, 548A, 2001.

49. J. C. Lindon, E. Holmes, J. K. Nicholson, *Prog. Nucl. Magn. Reson. Spectrosc.* 39, 1, 2001.

50. W. Windig, B. Antalek, *Chemom. Intell. Lab. Syst.* 46, 207, 1999.

https://biblioteca-digitala.ro / https://unibuc.ro

7. Procesarea semnalelor analitice

Prelucrarea semnalelor analitice reprezintă o parte fundamentală a chemometriei care are ca scop extragerea informației conținute în datele chimice obținute experimental prin investigarea sistemelor materiale de interes.

Din punct de vedere istoric, prin ***semnal*** se înțelege *orice set de semne sau simboluri capabile de a transmite informații sau mesaje* [1]. De aceea, sub denumirea de „semnale analitice" pot fi grupate toate tipurile de date obținute în practica analitică. Totuși, literatura analitică publicată sub denumirea generică de „Signal Processing" include doar tehnicile matematice de analiză a datelor înregistrate continuu, ca funcții care depind de una sau de mai multe variabile continue. În acord cu aceasta, prin semnal se va înțelege, în acest capitol, doar mulțimi de date cum sunt cromatogramele, spectrele, înregistrările obținute cu ajutorul analizoarelor continue sau cu instrumente care operează continuu etc.

7.1. Tipuri de semnale

Majoritatea semnalelor de interes științific sunt cuantificabile. Prin domeniul unui semnal se înțeleg variabilele independente în funcție de care se face analiza fenomenului investigat. Astfel, se pot menționa domeniile timpului (întâlnit la achiziția semnalelor de către analizoarele continue și cele cromatografice), domeniul spațial în două sau trei dimensiuni (cu aplicații numeroase la analiza imaginilor), domeniul lungimilor sau al numerelor de undă (cu aplicații în spectrometrie), domeniul potențialelor (cu aplicații în voltametrie) etc.

Instrumentele analitice au fost clasificate după numărul variabilelor de care depind semnalele produse de către acestea [2]. Astfel, analizoarele cromatografice și spectrometrele de diferite tipuri au fost denumite instrumente de ordinul întâi deoarece semnalele produse de acestea depind de o singură variabilă (timpul, lungimea de undă, raportul masă/sarcină etc). Spre deosebire de acestea, instrumentele de ordinul doi, rezultate prin cuplarea în tandem a două instrumente de ordinul întâi, furnizează semnale care admit o reprezentare tridimensională, acestea fiind funcții de două variabile.

https://biblioteca-digitala.ro / https://unibuc.ro

Cele mai reprezentative tehnici analitice cuplate sunt GC/MS, MS/MS, ^{2D}NMR şi GC/FT-IR. De asemenea, tot din această categorie fac parte seturile de sensori (multidiode, electrozi amperometrici în serie polarizaţi diferit etc) care au început să fie utilizate frecvent ca multidetectori în sistemele continue de analiză. Un exemplu de semnal tridimensional este prezentat în figura 7.1.

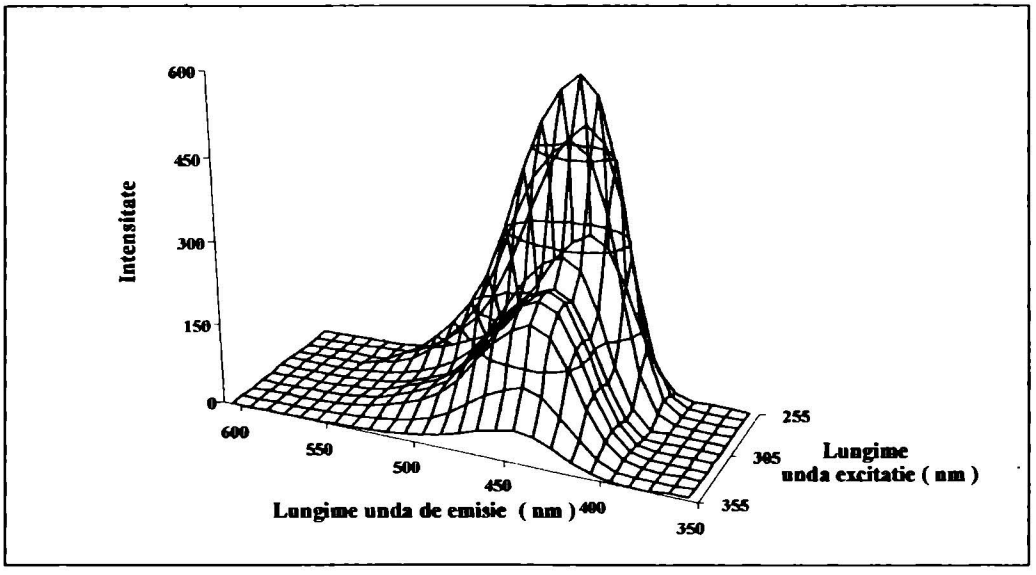


Figura 7.1. Spectrul de excitaţie-fluorescenţă al unei soluţii de ciprofloxacină.

Până în prezent, instrumentele de ordin trei, care furnizează semnale mai greu de vizualizat, sunt relativ puţine şi se întrebuinţează mai rar. Exemple de acest tip sunt măsurătorile de luminiscenţă în timp, în care se înregistrează intensitatea radiaţiei de luminiscenţă în funţie de timp şi de lungimile de undă ale acesteia şi ale radiaţiei excitatoare („***Time-resolved Luminescence***"). Avantajele utilizării acestor instrumente constau în posibilitatea de determinare exactă a componenţilor analizaţi din probe care conţin interferenţi ce nu au fost prezenţi în etaloanele utilizate la calibrare [2].

În continuare, vor fi prezentate doar câteva tehnici de prelucrare a semnalelor monovariabile obţinute cu instrumente de ordinul întâi. Acestea se aplică, în manieră corespunzătoare, şi semnalelor multidimensionale. De asemenea, deşi la studiul sem-

https://biblioteca-digitala.ro / https://unibuc.ro

nalelor monovariabile se utilizează de obicei domeniul timpului, teoriile generale obținute astfel se aplică și pentru semnale înregistrate în alte domenii.

La rândul lor, semnalele monovariabile pot fi clasificate după mai multe criterii. După valorile luate de variabila dependentă, pot fi menționate: semnale continue (temperatura etc), discontinue (numărul vânzărilor dintr-un produs) și fără amplitudine. Acestea din urmă dau informații despre timpul sau locul de desfășurare al unui fenomen și se numesc procese de punct. Ele au aplicații în codificarea neuronală a informației. De asemenea, după valorile luate de variabila independentă, se pot distinge: semnale în domeniul continuu, notate $x(t)$, sau în domeniu discret, notate $x(n)$ (vezi secțiunea următoare).

După formă, se disting semnale periodice, care au proprietatea $x(t) = x(t+P)$, unde P este durata unei perioade, și aperiodice. După durată: infinite sau finite.

Dintre toate tipurile de semnale, cel mai ușor de descris matematic sunt semnalele periodice cu durată infinită, care descriu procese ciclice relativ numeroase. Ca urmare a acestui avantaj, semnalele cu durată finită sunt considerate, frecvent, ca reprezentând o singură perioadă a unor semnale periodice infinite.

După natura proceselor descrise, se disting semnale ***deterministe*** și, respectiv, ***stochastice***. Acestea din urmă descriu procesele aleatoare care determină apariția zgomotului de fond conținut întotdeauna în semnalele reale în timp ce semnalele deterministe constituie componenta repatabilă a înregistrărilor obținute la desfășurarea în condiții identice a unui proces.

O altă clasificare a semnalelor poate fi efectuată după valoarea energiei acestora, E. Pentru semnale cu durată infinită în domeniu de timp continuu, ***energia*** se definește prin relația:

$$E = \int_{-\infty}^{+\infty} x^2(t)dt$$

, iar pentru cele din domeniu discret:

$$E = T\sum_{n=-\infty}^{+\infty} x^2(nT)$$

, în care T reprezintă intervalul de prelevare.

https://biblioteca-digitala.ro / https://unibuc.ro

Semnalele energetice sunt semnale cu energie finită. Astfel de semnale sunt cele tranziente, care prezintă valori diferite de zero într-un interval de timp finit. Alte tipuri de semnale, denumite semnale de putere, au energie infinită. Pentru acestea, se defineşte ***puterea***, P, prin:

$$P = \lim_{T_{tot} \to \infty} \frac{1}{2T_{tot}} \int_{-T_{tot}}^{+T_{tot}} x^2(t)dt$$

în cazul semnalelor cu durată infinită desfăşurate în domeniu continuu, şi respectiv:

$$P = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{+N} x^2(nT)$$

pentru cele în domeniu discret. În penultima ecuaţie, T_{tot} reprezintă jumătate din perioada totală a semnalului continuu.

7.2. Eşantionarea semnalelor

Semnalele analitice se desfăşoară de obicei în domeniul de timp continuu dar pentru prelucrarea lor cu ajutorul calculatoarelor, valorile se achiziţionează direct de către acestea. Calculatoarele nu admit însă decât o reprezentare în domeniu de timp discret a semnalelor, fiind mai puţin apte de a descrie mulţimi continue de valori. Procesul de transformare a semnalului continuu într-unul discret se realizează printr-o operaţie de conversie analogic/digitală (A/D) şi se numeşte ***eşantionare***. Aceasta este reprezentată schematic în figura 7.2. Semnalul $f(t)$ se măsoară continuu de către un sensor prevăzut cu un traductor. Traductorul transformă semnalul, de obicei proporţional, într-o mărime electrică, $x(t)$, care este apoi aplicată eşantionatorului reprezentat printr-un comutator cu perioada T şi având durata de trecere a semnalului τ. Eşantionatorul converteşte semnalul electric continuu într-o secvenţă de timp discret, $\big(x(T), x(2T), ..., x(nT), ..., x(NT)\big)$, denumită ***serie de timp***. Aceasta conţine valorile semnalului la anumite valori ale timpului care sunt separarte între ele prin aceeaşi durată, T, mai numită şi interval de prelevare, şi se reprezintă simplu astfel: $x(n), 0 \le n \le N-1$, prin omiterea intervalului de prelevare, care se presupune subînţeles.

https://biblioteca-digitala.ro / https://unibuc.ro

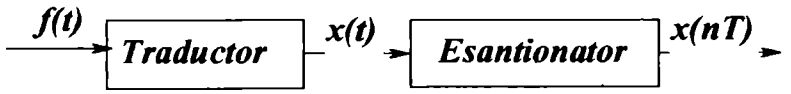


Figura 7.2. Reprezentarea schematică a procesului de eşantionare.

După modul de efectuare al conversiei A/D, în literatură [3] sunt menţionate dispozitivul de eşantionare mediator, care calculează valoarea medie a semnalului $x(t)$ pe durata τ a impulsului în acord cu relaţia:

$$x(n) = \frac{1}{\tau} \int_{nT-\tau}^{nT} x(t)dt$$

, şi dispozitivul cu autoblocare, care reţine valoarea instantanee a semnalului la momentul nT pe durata τ necesară efectuării conversiei.

Operaţia de eşantionare trebuie să conserve spectrul semnalului. Teorema de eşantionare a lui Shannon arată că un semnal poate fi reconstituit din eşantioanele sale cu condiţia ca *frecvenţa de eşantionare*, definită ca numărul de valori prelevate într-o perioadă de timp, să fie egală sau mai mare decât dublul celei mai înalte frecvenţe conţinută în spectrul semnal.

Convertoarele A/D tipice lucrează la frecveţe de eşantionare de ordinul a 100.000 de prelevări/secundă.

Operaţia de eşantionare produce nu doar discretizarea domeniului variabilei independente, timpul, ci şi a domeniului semnalului ca urmare a faptului că memoria calculatorului nu poate reprezenta o mulţime continuă şi, prin urmare, infinită de valori. Într-adevăr, memoria fiind compusă din cuvinte în cod binar cu o lungime finită, notată L, nu pot fi reprezentate distinct decât 2^L valori. De aceea, domeniul semnalului, $(x_{\min}, x_{\max}]$, este divizat de convertorul A/D în intervale de clasă cu lăţimea Δx de forma $(x_{min}+(i-1)\Delta x, x_{min}+i\Delta x]$, unde i sunt numere întregi variind de la 1 la 2^L. Toate valorile semnalului dintr-un anumit interval sunt aproximate prin mijlocul intervalului respectiv fiind reprezentate în memorie prin acelaşi cuvânt. În consecinţă, fie-

https://biblioteca-digitala.ro / https://unibuc.ro

care valoare a semnalului, $x(n)$, conține în sine o eroare de rotunjire, $\eta(n)$. Aceste erori sunt distribuite uniform în intervalul $\left(-\frac{\Delta x}{2}, \frac{\Delta x}{2}\right]$ și au varianța, σ^2, dată de relația:

$$\sigma^2 = \frac{(\Delta x)^2}{12}$$

, după cum se poate simplu verifica.

Așadar, pe lângă alte tipuri de zgomot conținute în semnal, operația de eșantionare introduce un zgomot suplimentar, denumit zgomot de cuantificare. Acesta poate fi minimizat prin reprezentarea în memorie a valorilor semnalului prin cuvinte cu lungime mai mare deoarece se micșorează astfel erorile de rotunjire prin reducerea lățimii intervalelor de calsă, care este dată de relația:

$$\Delta x = \frac{x_{max} - x_{min}}{2^L}$$

În practică, pentru majoritatea aplicațiilor, o lungime de 12 biți a cuvintelor care stochează valorile semnalelor este suficientă pentru neglijarea zgomotului de cuantificare. Această lungime variază, de obicei, între 4 și 32 de biți. Rareori se utilizează convertoare cu lungime a cuvintelor mai mare, doar în cazurile în care valorile semnalelor se găsesc într-un domeniu foarte larg și este necesară o precizie deosebit de bună. Totuși, în coprocesoarele matematice, care efectuează prelucrări complexe, lungimea cuvintelor poate depăși chiar 80 de biți.

7.3. Prelucrarea semnalelor analitice

Tehnicile matematice utilizate la prelucrarea semnalelor analitice sunt foarte diverse și, dintre acestea, se menționează aici doar câteva: medierea ansamblului, derivarea, transformarea, filtrarea și modularea semnalelor, testarea statistică de ipoteze, modelarea statistică a componentei stochastice a semnalelor, analiza spectrală, convoluții și deconvoluții, interpolări și extrapolări, precum și tehnicile de regresie. Principiile unora dintre aceste tehnici se desprind din capitolele precedente.

Tehnicile de prelucrare a semnalelor pot fi aplicate „*on-line*" de către instrumente specializate sau pot fi implementate, utilizând produse software adecvate, pe calculatoare cuplate cu instrumentele analitice. Aceste prelucrări pot fi efectuate în

https://biblioteca-digitala.ro / https://unibuc.ro

timp real sau „*off-line*". Pentru a uşura înţelegerea modalităţilor de aplicare a acestora în analiza chimică, vor fi prezentate mai întâi, într-o manieră succintă şi generală, instrumentele matematice de bază utilizate cel mai frecvent la prelucrarea seriilor de timp.

7.3.1. Aspecte fundamentale ale prelucrării semnalelor

Spre deosebire de prelucrarea altor tipuri de date (cum sunt, de exemplu, valorile absorbanţelor unor etaloane utilizate pentru a calcula ecuaţii de calibrare), în analiza matematică a seriilor de timp trebuie acordată o importanţă deosebită ordinei în care sunt colectate valorile semnalelor deoarece, în majoritatea cazurilor, acestea depind de valorile anterioare. Astfel de situaţii se modelează matematic cu ajutorul lanţurilor stochastice de tip Markov şi apar la înregistrarea oricăror semnale deoarece instrumentele de măsură prezintă o inerţie de răspuns care se traduce printr-o constantă de timp diferită de zero.

Cu privire la prelucrarea semnalelor, trebuie avute în vedere următoarele două aspecte fundamentale: ***caracterul*** dublu, ***temporal*** şi ***frecvenţial***, al semnalelor şi ***contaminarea cu zgomot*** a semnalelor experimentale.

7.3.1.1. Caracterul temporal – frecvenţial al semnalelor

Fiecare semnal achiziţionat în domeniul timpului poate fi considerat ca fiind compus dintr-o sumă de unde sinusoidale sau cosinusoidale denumite componente spectrale iar determinarea parametrilor acestora (amplitudine – mai denumită uneori şi magnitudine, frecvenţă şi fază) se numeşte analiză spectrală.

În mod paradoxal, deşi majoritatea semnalelor sunt obţinute în domeniul timpului de către instrumentele analitice, deseori acestea sunt mai sugestive în domeniul frecvenţelor. De asemenea, chiar şi în natură, percepţia senzorială (auzul, văzul) se face în domeniul frecvenţelor.

Reprezentările semnalelor în domeniile timpului şi al frecvenţelor sunt complementare. Astfel, în timp ce domeniul timpului asigură informaţii despre amplitudinea semnalului la diverse momente de timp, domeniul frecvenţelor conţine informaţii despre frecvenţele prezente în semnal şi magnitudinile lor.

https://biblioteca-digitala.ro / https://unibuc.ro

Modul de efectuare al analizei spectrale depinde de tipul semnalelor. Pentru semnale deterministe desfăşurate în domeniu de timp continuu se aplică în acest scop fie dezvoltarea în serii Fourier (dacă semnalul este periodic), fie transformata Fourier continuă (pentru semnalele aperiodice). Acestea sunt prezentate în anexa 1.

Întrucât calculatoarele electronice nu operează cu funcţii continue, în practică se utilizează, pentru analiza spectrală a seriilor de timp, varianta discretă a transformatei Fourier, notată DFT, în care atât domeniul timpului, cât şi cel al frecvenţelor au valori discrete. Pentru aceasta, este însă necesar ca numărul valorilor din seria de timp, N, să fie finit.

DFT este un caz particular al următoarei ecuaţii generale de transformare a seriilor de timp:

$$X(m) = \sum_{n=0}^{N-1} x(n) K(m,n)$$

, în care $K(m,n)$ reprezintă factorul de transformare aplicat seriei de timp. Pentru DFT, acesta are expresia:

$$X(m,n) = \exp(-j2\pi mn / N)$$

Considerăm că un semnal continuu, $x(t)$, cu durata P, ($P=NT$), a fost eşantionat cu un interval de prelevare T, obţinându-se seria de timp $x(n), 0 \le n \le N-1$. În vederea calculării transformatei Fourier discrete a unui astfel de semnal, domeniul frecvenţelor se discretizează în N valori, $(f_m) \cdot m = 0, N-1$, fiecare dintre acestea reprezentând un multiplu întreg al frecvenţei de spaţiere, f_s:

$$f_m = mf_s = m\frac{1}{P}$$

Una din definiţiile perechilor de transformate Fourier discrete [1] este reprezentată prin ecuaţiile:

$$X(m) = \frac{1}{N}\sum_{n=0}^{N-1} x(n) \exp\left(-\frac{j2\pi mn}{N}\right), 0 < m < N-1$$

si:

$$x(n) = \sum_{m=0}^{N-1} X(m) \exp\left(\frac{j2\pi mn}{N}\right), 0 < n < N-1$$

https://biblioteca-digitala.ro / https://unibuc.ro

În primele din aceste ecuații, prin $X(m)$ s-a notat valoarea transformatei Fourier directe în punctul $f = f_m$ iar ultimele ecuații, care permit calculul semnalului în domeniul de timp discret cunoscând caracteristicile de frecvență ale acestuia, exprimă transformata Fourier inversă.

Se observă că efectul discretizării domeniului frecvențelor constă în transformarea semnalului inițial finit într-unul infinit și periodic cu perioada P, ceea ce este echivalent cu: $x(n) = x(n + kN)$, pentru $k \in Z$.

Calculul transformatei Fourier discrete cu ecuațiile de mai sus este foarte laborios deoarece fiecare set de ecuații implică un număr de N^2 operații de multiplicare a unor numere complexe, ceea ce, pentru N mare, poate necesita un timp de calcul foarte lung chiar dacă acesta se efectuează cu calculatoare performante. Algoritmul Cooley-Tukey (1965) de calcul rapid al transformatei Fourier, a cărui descriere poate fi găsită în lucrarea [4], reduce semnificativ timpul de calcul deoarece numărul de operații necesare este mult mai mic, fiind egal cu N log 2. Astfel, aplicarea sa pentru $N = 1024$ este de cca 100 de ori mai rapidă. Singura condiție necesară în acest algoritm este ca numărul valorilor din seria de timp să fie o putere naturală a lui 2 ($N = 2^l$).

În general, numărul valorilor din semnal nu prezintă restricții. Cu toate acestea, considerarea unor valori pare pentru N prezintă unele avantaje în interpretarea DFT după cum va rezulta în continuare. De aceea, mai departe se presupune îndeplinită această condiție.

Cunoașterea detaliată a proprietăților DFT este necesară pentru a minimiza numeroasele tipuri de erori care pot să apară la calcularea sa și deoarece multe tehnici de prelucrare a semnalelor analitice sunt bazate pe ele. Aceste proprietăți pot fi rezumate prin următoarele teoreme ale căror demonstrații sunt imediate:

Liniaritatea. DFT a unei combinații liniare a două semnale, $x(n)$ și $y(n)$, prelevate la aceleași valori ale timpului, este egală cu aceeași combinație liniară a transformatelor Fourier corespunzătoare sau, în formă simbolică, dacă $x(n) \Leftrightarrow X(m)$ și $y(n) \Leftrightarrow Y(n)$ atunci $ax(n) + by(n) \Leftrightarrow aX(m) + bY(m)$, pentru orice valori complexe ale lui a și b.

https://biblioteca-digitala.ro / https://unibuc.ro

Periodicitatea. DFT a oricărei serii de timp care se produce într-un domeniu de timp finit este o funcție periodică cu perioada egală cu Nf_s, ceea ce se exprimă prin:

$$X(m) = X(m + kN), \text{ pentru } k \in Z$$

Simetria conjugată. În domeniul frecvențelor, componentele simetrice ale transformatei Fourier discrete ale oricărui semnal, $x(n)$, reprezintă valori complex conjugate, adică pentru orice m $(1 < m < N/2 - 1)$:

$$X(N - m) = X^*(m)$$

Teorema de deplasare în domeniul timpului. Dacă $x(n)$ şi $X(m)$ sunt două perechi de transformate Fourier discrete, atunci transformata Fourier directă a unui semnal $y(n) = x(n\text{-}k)$ se calculează cu relația:

$$Y(m) = X(m)\exp\left(-\frac{j2\pi mk}{N}\right)$$

Convoluția în domeniul frecvenței. Fiind date semnalele $x(n)$ şi $y(n)$ prelevate la aceleaşi valori ale timpului şi transformatele lor Fourier, $X(m)$ şi $Y(m)$, transformata Fourier discretă a produsului celor două semnale se calculează cu relația:

$$\text{DFT}[x(n)y(n)] = \sum_{k=0}^{N-1} Y(k)X(m-k)$$

si se notează $X(m)*Y(m)$.

Convoluția în domeniul timpului. Fiind date semnalele frecvențiale $X(m)$ şi $Y(m)$ cunoscute la aceleaşi frecvențe şi transformatele lor Fourier, $x(n)$ şi $y(n)$, transformata Fourier discretă a produsului semnalelor frecvențiale se calculează cu relația următoare:

$$\text{DFT}[X(m)Y(m)] = \frac{1}{N}\sum_{k=0}^{N-1} x(k)y(n-k)$$

si se notează $x(n)*y(n)$.

Ultimele două teoreme pot fi reprezentate simbolic astfel: $\text{DFT}[x(n)y(n)] \Leftrightarrow$ $\Leftrightarrow X(m)*Y(m)$, şi $\text{DFT}[X(m)Y(m)] \Leftrightarrow x(n)*y(n)$. Sensul lor poate fi rezumat spunând că procesul de convoluție a două semnale într-un domeniu echivalează cu produsul scalar al semnalelor în domeniul transformat. Ultimele relații se numesc ***ecuații de convoluție*** în domeniul frecvențelor şi, respectiv, al timpului.

https://biblioteca-digitala.ro / https://unibuc.ro

Teoremele de mai sus prezintă câteva consecințe importante. Astfel, teorema de simetrie conjugată arată că, dintre cele N componente ale DFT[$X(N)$], doar $N/2+1$ sunt independente, şi anume $X(m)$ pentru $0<m<N/2$, motiv pentru care unele programe de calcul furnizează doar valorile acestora. Componentele pentru $N/2+1<m<N-1$, exprimă o informație redundantă deoarece acestea se confundă cu valorile $X(m)$ pentru $1<m<N/2-1$. În continuare, efectul acesta, („*folding*" în limba engleză), îl vom denumi ***redundanță***.

Componentele $X(N/2)$ şi $X(0)$ au întotdeauna valori reale iar ultima dintre acestea, ***termenul DC***, este egală cu media valorilor semnalului din domeniul timpului.

Datorită redundanței, frecvențele f_m pentru $N/2+1<m<N-1$ nu se deosebesc de cele pentru $1<m<N-1$. Pentru a dovedi acest lucru, observăm că, datorită periodicității şi simetriei conjugate, au loc relațiile:

$$X(N-m)=X(-m)=X^*(m)$$

pentru fiecare m, $0<m<N/2-1$. Ultima parte a ecuațiilor precedente este similară cu ecuația A10 şi cu relația dintre coeficienții seriilor Fourier. Ecuațiile de calcul a transformatei Fourier inverse pot fi rescrise astfel:

$$x(n)=\sum_{m=-N/2+1}^{N/2} X(m)\exp\left(\frac{j2\pi mn}{N}\right),\ 0<n<N-1$$

si, mai departe, prin analogie cu seriile Fourier (vezi anexa 1):

$$x(n)=C_0\sum_{m=1}^{N/2} C_m\cos(2\pi f_m nT+\theta_m),\ 0<n<N-1$$

, unde $C_0=X(0)$; $C_m=|X(m)|$; $\theta_m=-\arctan\dfrac{\mathrm{Im}[X(m)]}{\mathrm{Re}[X(m)]}$, iar restul mărimilor au semnificațiile cunoscute. Ultimele ecuații dovedesc prezența în spectrul semnalului obținut prin aplicarea DFT a numai $N/2$ frecvențe discrete, în afara termenului DC.

Cea mai mare frecvență neredundantă din semnal apare pentru $m=N/2$. Ea este jumătate din frecvența de prelevare şi se numeşte ***frecvență Nyquist***. DFT este definită în mod unic doar pentru semnale conținând frecvențe mai mici decât acestă frecvență şi, deci, având un spectru de bandă limitat, fapt concordant cu teorema de eşantionare a lui Shannon.

https://biblioteca-digitala.ro / https://unibuc.ro

Spectrul din regiunea frecvențelor negative este complexul conjugat al celui din regiunea pozitivă. Aceasta poate fi observat și în figura 7.3 în care sunt prezentate atât valorile reale și imaginare ale componentelor DFT, cât și spectrele de magnitudine și de fază corespunzătoare unui semnal gaussian care simulează un pic cromatografic. Se remarcă că semnalul analitic prezentat, care simulează componenta deterministă a unor semnale înregistrate în mod real, are cea mai mare parte a energiei repartizată pe un număr redus de frecvențe situate în vecinătatea frecvenței zero (vezi figura 7.3d).

O caracteristică a DFT constă în aceea că asigură valori ale lui $X(f)$ doar la anumite frecvențe. Aceasta nu are consecințe grave în cazul unui spectru continuu. Totuși, dacă o valoare importantă din spectru, cum ar fi, de exemplu, maximul unui pic foarte ascuțit, există la o frecvență $f \neq mf_s$, aceasta nu va putea fi pusă în evidență. Această eroare („*picket fence effect*") se numește ***efect de grilă***. Ea se poate înlătura fie prin mărirea frecvenței de achiziție a semnalului în aceeași durată de timp, fapt nu întotdeauna posibil de realizat, fie prin completarea semnalului cu un anumit număr de zerouri. Cea de a doua metodă constituie, de fapt, o metodă de interpolare bazată pe extinderea arbitrară a duratei semnalului.

Dacă la sfârșitul unui semnal cu durata NT se adaugă M zerouri, se obține un semnal nou cu durata $LT = (M + N)T$. La calculul DFT al noului semnal, frecvența de spațiere, f_s, dată de relația:

$$f_s' = \frac{1}{LT} = \frac{1}{(M+N)T}$$

se va micșora, ceea ce va permite evidențierea unui număr mai mare de frecvențe.

Completarea cu zerouri nu modifică forma de ansamblu a spectrului. Ea se aplică și la utilizarea algoritmilor de calcul rapid al DFT în cazurile în care numărul valorilor din semnale diferă de o putere naturală a lui 2.

O sursă majoră de erori la calcularea DFT apare ca o consecință a trunchierii semnalelor care este întotdeauna prezentă la prelevarea unui semnal cu durată finită dintr-un semnal cu durată teoretic infinită.

https://biblioteca-digitala.ro / https://unibuc.ro

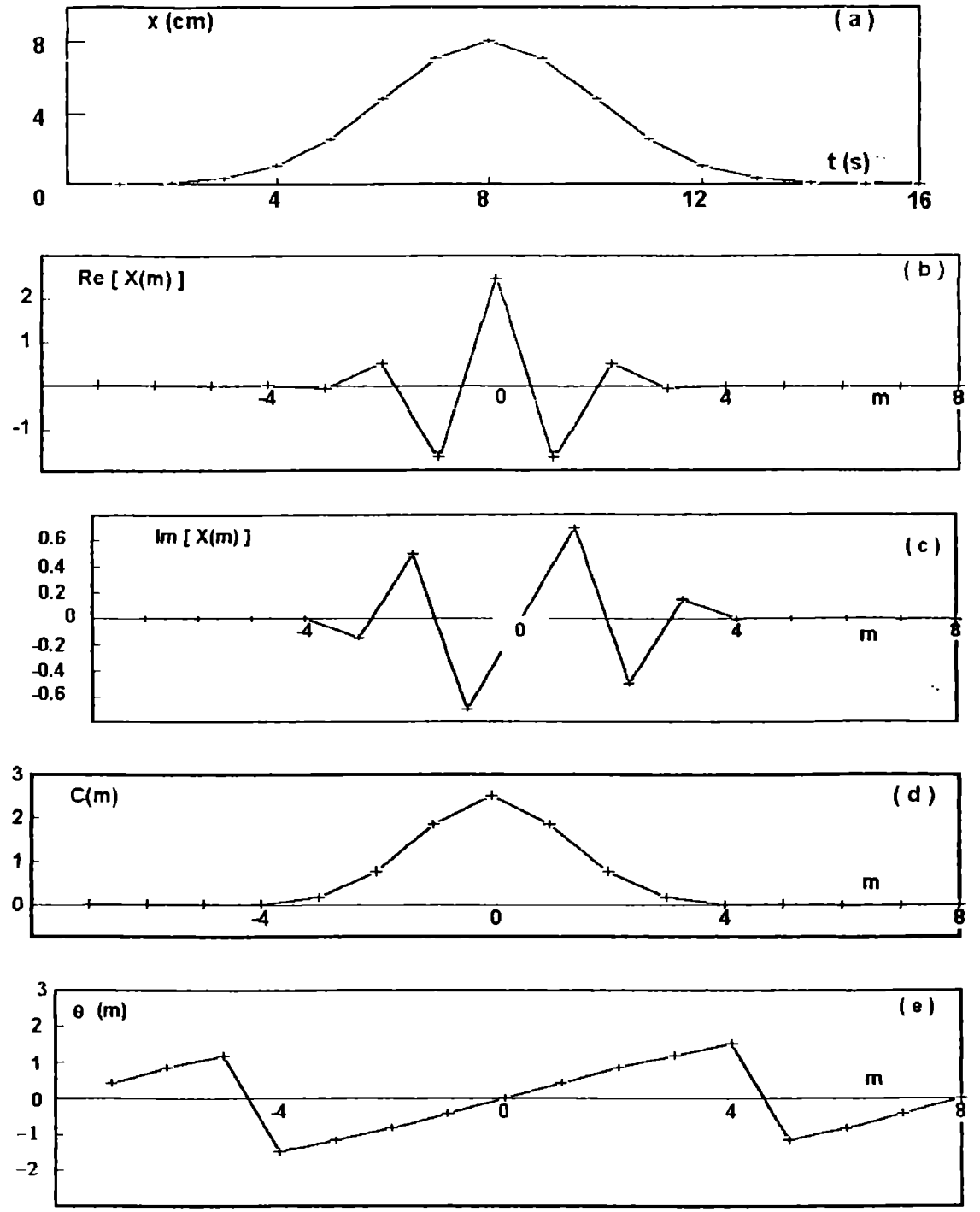


Figura 7.3. Semnal discret (a) obținut prin eşantionarea (N=16, T=1 s) unui pic cromatografic simulat având următorii parametri: înălțimea maximă – 8 cm; poziția maximului – 8 s; varianța – 4 s^2; Valorile reale (b) şi imaginare (c) ale spectrului semnalului; Spectrele de magnitudine (d) şi de fază (e) ale semnalului.

Trunchierea poate fi descrisă matematic considerând că semnalul real, $x(n)$, se multiplică cu o fereastră de date dreptunghiulară, $d_R(n)$, pentru a obține semnalul prelevat, $y(n)$:

https://biblioteca-digitala.ro / https://unibuc.ro

$$y(n) = x(n)d_R(n)$$

, unde: $d_R(n) = \begin{cases} 1, \text{ pentru } 0 < n < N-1 \\ 0, \text{ pentru alti } n \end{cases}$

DFT[$y(n)$] diferă de DFT[$x(n)$] datorită teoremei de convoluție:

$$Y(m) = X(m) * D_R(m) \qquad (7.1)$$

, unde: $y(m) \Leftrightarrow Y(m)$; $x(n) \Leftrightarrow X(m)$ si $d_R(n) \Leftrightarrow D_R(m)$. Genul acesta de eroare („*leakage error*") se numeşte ***eroare de dispersare***.

Pentru a vedea care sunt consecințele erorii de dispersare, considerăm două semnale în domeniul continuu; [$x_i(t)$] (i = 1, 2), definite prin:

$$x_i(t) = \cos(2\pi f_{i0} t)$$

, cu f_{10}=1/8 si f_{20}=1/9. Spectrele Fourier ale acestora, obținute prin aplicarea CFT, sunt reprezentate în figura 7.4 şi au expresia:

$$X_i(f) = 0{,}5\delta(f - f_{i0}) + 0{,}5\delta(f + f_{i0}),\ i = 1{,}2$$

în care δ semnifică funcția lui Dirac.

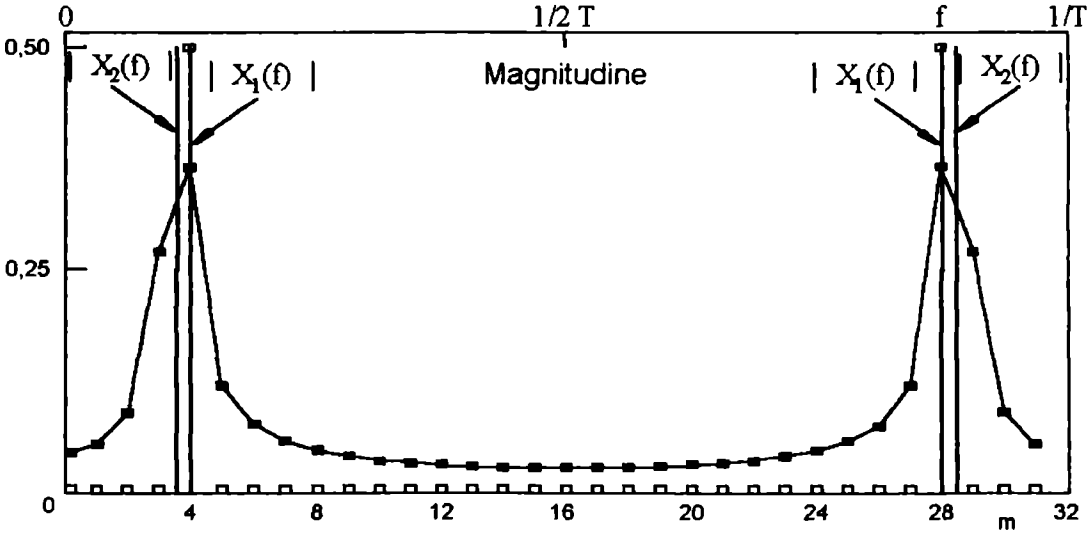


Figura 7.4. Spectrele Fourier de magnitudine continue, $|X_i(f)|$, şi discrete, $|X_i(m)|$, pentru cele două semnale sinusoidale definite în text. ▫ pentru $|X_1(m)|$ si ▪ pentru $|X_2(m)|$.

https://biblioteca-digitala.ro / https://unibuc.ro

Considerăm că semnalele $x_i(t)$ sunt discretizate, începând cu momentul $t = 0$, în N=32 de valori cu o frecvență de eşantionare f_e=1/NT=1/32, obținându-se semnalele $x_i(N)$, $0 \le N \le 31$. Transformatele Fourier discrete ale acestor două semnale, $X_i(m)$, sunt de asemenea prezentate în figura 7.4. Se observă că $X_1(m)$ prezintă, pentru m=4 şi m=28, valorile 0,5, ceea ce este corect, semnificând prezența în semnal a frecvenței f= $4f_e$ =1/8 cu magnitudinea 1 (2 x 0,5). Pentru al doilea semnal însă, deoarece $f_{20} \neq mf_e$, apare o distorsionare importantă a spectrului magnitudinilor calculat aplicând DFT, ceea ce este o consecință directă a trunchierii semnalului.

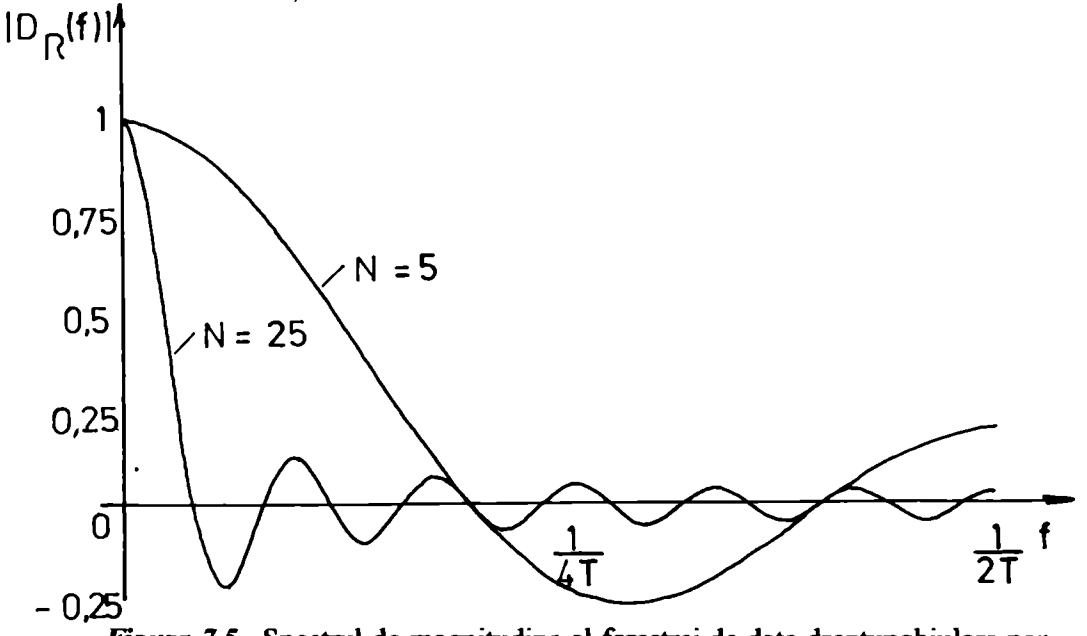


Figura 7.5. Spectrul de magnitudine al ferestrei de date dreptunghiulare pentru două valori diferite ale numărului valorilor din semnal.

Eroarea de dispersare este inacceptabilă. Minimizarea sa poate fi înțeleasă analizând forma factorului de convoluție, $D_R(m)$, din ecuația 7.1, mai denumit *fereastră spectrală dreptunghiulară*. În domeniul de frecvențe continuu, ecuația sa este:

$$D_R(f) = \frac{\sin(\pi f NT)}{\sin(\pi f T)} \exp(-j\pi f NT)$$

, iar spectrul său Fourier, care este o funcție pară, este prezentat în figura 7.5 pentru două valori diferite ale lui N doar în domeniul frecvențelor pozitive. Se observă că există multe intersectări ale axei frecvențelor iar porțiunile dintre două intersectări con-

https://biblioteca-digitala.ro / https://unibuc.ro

secutive se numesc ***lobi***. Dintre aceştia, lobul centrat în jurul frecvenţei zero reprezintă lobul principal iar ceilalţi – lobul secundar.

Operaţia de convoluţie dispersează efectul de ponderare al lobilor secundari asupra întregului domeniu de frecvenţe. Se poate dovedi că din ecuaţia 7.1 rezultă $X(m)=Y(m)$ doar dacă fereastra spectrală de convoluţie este identică cu funcţia lui Dirac care descrie un pic frecvenţial infinit îngust. Din figura 7.5 se observă că pentru N mai mare, $D_R(f)$ prezintă un lob principal mai apropiat de forma funcţiei Dirac fiind mai îngust. Aşadar, o modalitate de micşorare a erorii de dispersare constă în mărirea lui N, ceea ce însă nu este întotdeauna posibil. De aceea, în practică se recurge la o altă metodă, şi anume la modificarea formei ferestrei de date, $d(n)$, cu scopul de a obţine ferestre spectrale, $D(m)$, ale căror lobi secundari să conţină mult mai puţină energie decât lobul principal.

În literatură au fost propuse numeroase ferestre de date diferite. Dintre acestea, figura 7.6 prezintă doar fereastra dreptunghiulară deja discutată şi fereastra triunghiulară a lui Bartlett împreună cu ferestrele spectrale corespunzătoare pentru o aceeaşi valoare a lui N.

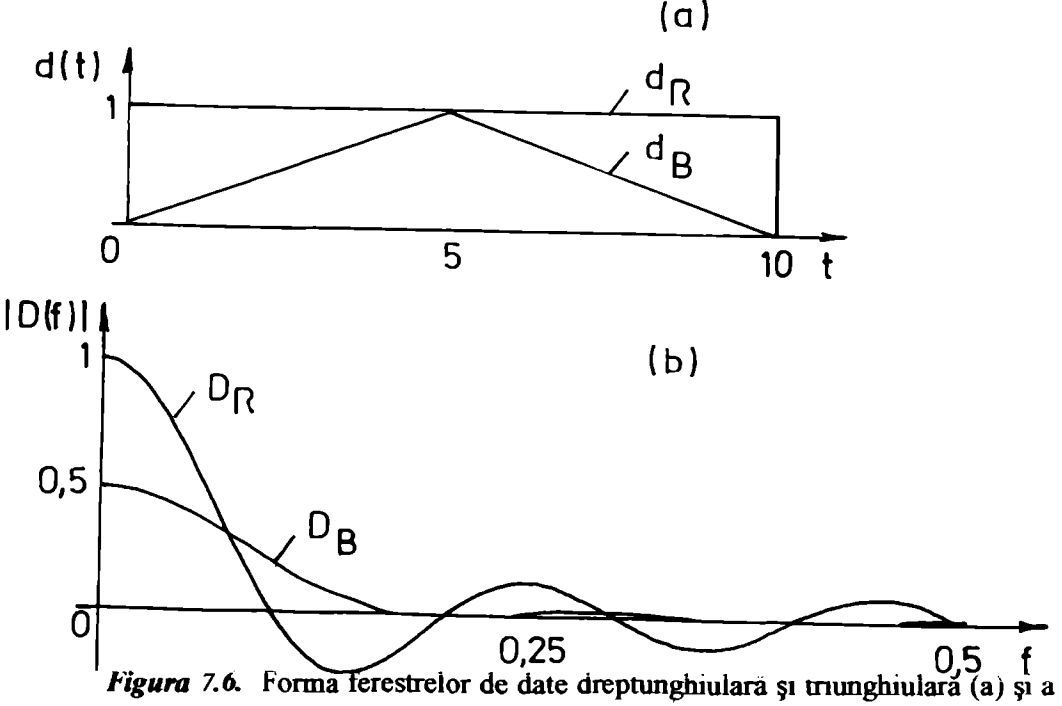


Figura 7.6. Forma ferestrelor de date dreptunghiulară şi triunghiulară (a) şi a ferestrelor spectrale corespunzătoare (b).

https://biblioteca-digitala.ro / https://unibuc.ro

Caracteristicile principale care diferențiază ferestrele spectrale sunt nivelul lobului principal, definit ca raportul dintre magnitudinile lobului principal şi al celui mai mare lob secundar, şi lățimea lobului principal, care reprezintă lățimea domeniului de frecvențe al acestuia. Aceste caracteristici, împreună cu ecuațiile ferestrelor de date spectrale corespunzătoare, sunt prezentate în tabelul 7.1.

Tabelul 7.1

Două ferestre de date şi caracteristicile ferestrelor spectrale corespunzătoare

Fereastra de date d(n)	Fereastra spectrală	Lățimea lobului principal	Nivelul lobului principal
Dreptunghiulară $d_R(n) = 1, 0 \le n \le N-1$	$D_R(f) = T\frac{\sin(\pi fNT)}{\sin(\pi fT)} x$ $x \exp[-j\pi T(N-1)]$	$\frac{2}{NT}$	22,4%
Triunghiulară-Bartlett $d_B(n) = \begin{cases} \frac{2n}{N-1}, \\ 0 \le n \le (N-1)/2 \\ 2 - \frac{2n}{N-1}, \\ (N-1)/2 \le n \le N-1 \end{cases}$	$D_B(f) = \frac{2T}{N}\left(\frac{\sin(\pi fNT/2}{\sin(\pi fT)}\right)^2$	$\frac{4}{NT}$	4,5%

Atât din tabelul 7.1, cât şi din figura 7.6 rezultă că micşorarea nivelului lobului principal este însoțită de creşterea lățimii acestuia, ceea ce are ca efect scăderea rezoluției. Într-adevăr, simulările pe calculator [1] ale DFT pentru un semnal de forma:

$$x(n) = \cos(2\pi f_1 nT) + \cos[2\pi(f_1 + \Delta f)nT]$$

, utilizând diferite ferestre de date, arată că cele două frecvențe conținute în semnal, f_1 şi $f_1+\Delta f$, sunt evidențiate distinct în spectrul Fourier doar dacă f_1 şi Δf este mai mică decât jumătate din lățimea lobului principal al ferestrei spectrale corespunzătoare ferestrei de date utilizate.

Alegerea ferestrei de date optime depinde de scopul urmărit. În general, utilizarea unor ferestre asemănătoare cu cea triunghiulară face ca spectrul Fourier al semnalului prelevat să fie mult mai apropiat ca formă de cel al semnalului real. Totuşi, ca urmare a diminuării amplitudinilor valorilor semnalului din domeniul timpului, în astfel de cazuri se pierde o parte din energia semnalului, fenomenul acesta fiind denumit ***pierdere de procesare***. El poate fi corectat prin multiplicarea energiei calculate cu un factor care depinde de fereastra de date utilizată.

https://biblioteca-digitala.ro / https://unibuc.ro

7.3.1.2. Contaminarea cu zgomot a semnalelor reale

Interpretarea semnalelor analitice este mult îngreunată de faptul că înregistrările obținute la monitorizarea unor procese fizico-chimice reale se compun întotdeauna dintr-un semnal stochastic denumit zgomot suprapus peste semnalul determinist de interes. Tehnicile prin care se înlătură parțial consecințele negative ale contaminării cu zgomot a semnalelor reale asupra estimării corecte a parametrilor fizico-chimici de interes sunt fundamentate pe metode matematice care asimilează zgomotul cu procese aleatoare tratate într-o manieră specifică. Noțiuni privind modul de reprezentare și de caracterizare al proceselor aleatoare sunt prezentate succint în continuare cu scopul de a facilita înțelegerea proprietăților zgomotului din sistemele analitice de măsură și a tehnicilor de combatere ale acestuia.

Un ***proces aleator*** reprezintă o variabilă aleatoare dependentă de timp și poate fi descris printr-un semnal stochastic pe care îl vom nota $x(t)$ în domeniu de timp continuu și, respectiv, $x(n)$ în domeniu discret. Deși aceste notații sunt identice cu cele utilizate în mod uzual și pentru semnale deterministe, proprietățile celor două tipuri de semnale nu trebuie confundate. Diferențele majore provin din faptul că semnalele deterministe sunt repetabile pe când cele stochastice nu. De aceea, caracterizarea acestora din urmă nu este posibilă decât în termeni probabilistici, pe baza unui ansamblu format dintr-un număr mare de înregistrări a procesului aleator fiecare dintre acestea fiind denumită funcție de probă.

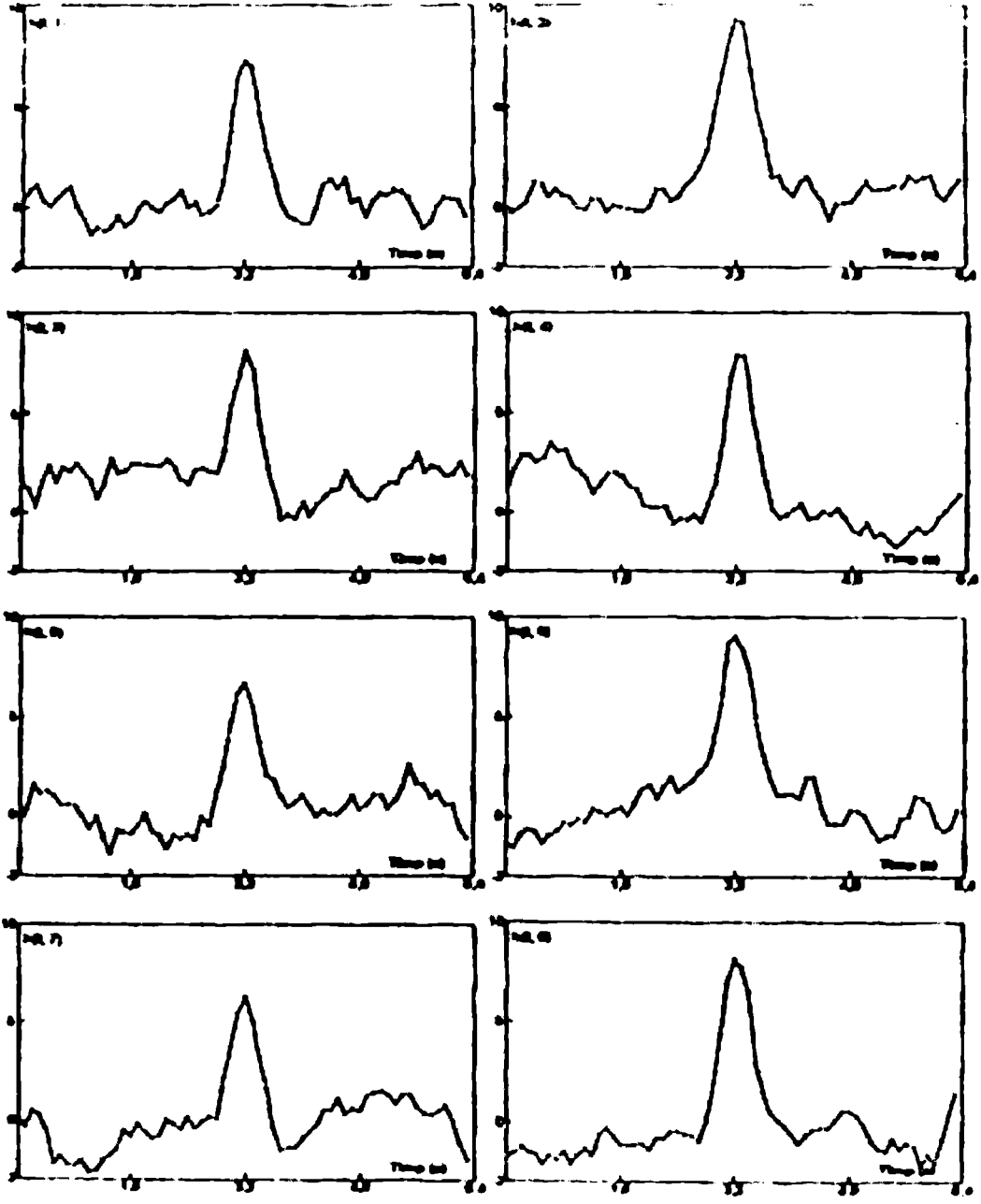
Ansamblul poate fi descris cu o funcție bivariabilă, $x(t, \zeta_n)$, în care argumentul cu valori discrete și naturale, ζ_n, reprezintă numărul înregistrării. În referința [5] sunt prezentate noțiuni de simulare a proceselor aleatoare necesare la testarea metodelor de prelucrare a semnalelor stochastice și la generarea de semnale cu proprietăți cunoscute iar în figura 7.7 este simulat un ansamblu format din 16 funcții de probă reprezentând un pic de formă gaussiană contaminat cu zgomot.

La fiecare valoare a timpului, t_0, valorile funcțiilor de probă, ($x(t_0,1)$, $x(t_0,2)$, $x(t_0,3)$, etc), reprezintă o variabilă aleatoare. Notând cu $F_{x(t0)}(\alpha)$ funcția de distribuție a acesteia, care reprezintă probabilitatea ca la momentul t_0 semnalul să aibă o valoare mai mică decât α, distribuția variabilei aleatoare, $x(t_0, \zeta_n)$, se caracterizează cu o fun-

https://biblioteca-digitala.ro / https://unibuc.ro

cție de probabilitate de ordinul unu, $f_{x(t_0)}(\alpha)$, definită similar cu cea pentru variabilele aleatoare obișnuite:

$$f_{x(t_0)}(\alpha) = \frac{dF_{x(t_0)}(\alpha)}{d\alpha}$$



https://biblioteca-digitala.ro / https://unibuc.ro

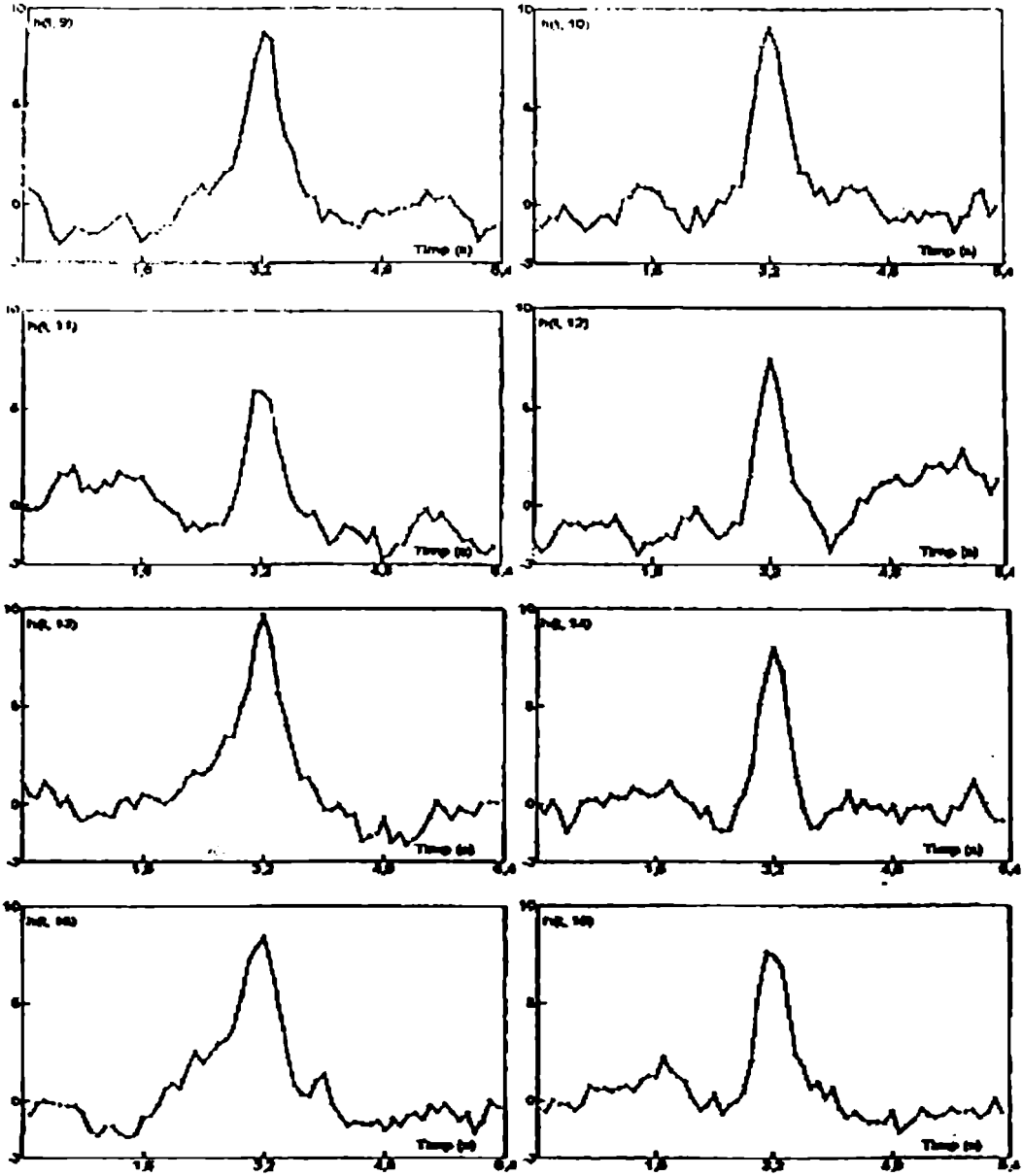


Figura 7.7. Picuri gaussiene simulate cu următoarele caracteristici: poziția maximului picului m=3,3 s; h_{max}=8 și σ = ,2 s, peste care s-a suprapus zgomot cu structură autoregresivă generat cu relația: $y(n)$=0,9 $y(n\text{-}1)$ + 0,5 $x(n)$. Raportul S/N = 6.

În cazul considerării a două momente de timp, t_1 și t_2, se definește funcția de probabilitate de ordinul doi, $f_{x(t_1)x(t_2)}(\alpha_1,\alpha_2)$, iar în cazul mai multor valori ale timpului – funcții de probabilitate de ordin superior.

https://biblioteca-digitala.ro / https://unibuc.ro

Utilizând funcțiile de probabilitate, cu relațiile următoare se definesc funcțiile de moment ale ansamblului:

$$m_{x(t_0)} = E[x(t_0)] = \int_{-\infty}^{+\infty} \alpha f_{x(t_0)}(\alpha) d\alpha$$

$$\sigma^2_{x(t_0)} = E\left[(x(t_0) - m_{x(t_0)})^2\right] = \int_{-\infty}^{+\infty} (\alpha - m_{x(t_0)})^2 f_{x(t_0)} d\alpha$$

$$\varphi_{x(t_1,t_2)} = E[x(t_1)x(t_2)] = \int_{-\infty-\infty}^{+\infty+\infty} \alpha_1 \alpha_2 f_{x(t_1)x(t_2)}(\alpha_1, \alpha_2) d\alpha_1 d\alpha_2$$

$$\gamma_{x(t_1,t_2)} = E[(x(t_1) - m_{x(t_1)})(x(t_2) - m_{x(t_2)})] =$$

$$= \int_{-\infty-\infty}^{+\infty+\infty} (\alpha_1 - m_{x(t_1)})(\alpha_2 - m_{x(t_2)}) f_{x(t_1)x(t_2)}(\alpha_1, \alpha_2) d\alpha_1 d\alpha_2$$

$$\xi_{x(t_1,t_2)} = \frac{\varphi_{x(t_1,t_2)} - m_{x(t_1)} m_{x(t_2)}}{\sigma_{x(t_1)} \sigma_{x(t_2)}} = \frac{\gamma_{x(t_1,t_2)}}{\sigma_{x(t_1)} \sigma_{x(t_2)}}$$

Media, $m_{x(t_0)}$, şi varianța, $\sigma^2{}_{x(t_0)}$, ansamblului reprezintă valorile de aşteptare (sau cele mai probabile) ale semnalului şi, respectiv, ale pătratului deviației semnalului de la medie la $t = t_0$. Ele sunt funcții de un singur moment de timp şi se numesc, de aceea, momente de ordinul întâi spre deosebire de funcțiile de autocorelație (ACF), de autocovarianță (ACVF) şi de autocovarianță normalizată (NACVF), notate $\varphi_{x(t_1,t_2)}$, $\gamma_{x(t_1,t_2)}$ şi, respectiv, $\xi_{x(t_1,t_2)}$, care depind de două momente de timp, t_1 şi t_2, şi sunt numite, de aceea, momente de ordinul doi. ACF este valoarea de aşteptare a produsului semnalelor la momentele t_1 şi t_2. ACVF semnifică valoarea de aşteptare a produsului deviațiilor de la medie a semnalului la momentele considerate iar NACVF reprezintă ACVF normalizată în raport cu produsul deviațiilor standard ale semnalului.

Funcțiile de moment reprezintă caracteristici fundamentale cu ajutorul cărora se descriu procesele aleatoare în domeniul timpului. Analiza semnalelor are ca scop determinarea acestora şi a funcțiilor de probabilitate. Utilizarea în acest scop a unui ansamblu format dintr-un număr mare de funcții de probă este practic nefezabilă. Este preferabil ca aceste determinări să se efectueze pe baza unei singure funcții de probă discretizate. Condițiile în care poate fi făcut acest lucru sunt rezumate în ***teoremele de***

https://biblioteca-digitala.ro / https://unibuc.ro

ergodicitate. Câteva concluzii ale acestora, de importanță fundamentală în prelucrarea semnalelor analitice, sunt prezentate sumar în continuare. Fiind vorba despre o singură funcție de probă din cadrul ansamblului, argumentul ζ_n va fi omis.

Un proces aleator prezintă staționaritate de ordinul întâi dacă funcția de probabilitate de ordinul unu nu depinde de timp ceea ce presupune respectarea relației:

$$f_{x(t_1)}(\alpha) = f_{x(t_2)}(\alpha)$$

, pentru oricare două momente de timp, t_1 și t_2, arbitrare iar staționaritatea de ordinul doi are loc dacă:

$$f_{x(t_1+\theta)x(t_2+\theta)}(\alpha_1,\alpha_2) = f_{x(t_1)x(t_2)}(\alpha_1,\alpha_2)$$

, pentru oricare t_1 și t_2 și pentru orice interval de timp, θ.

Procesele aleatoare reale prezintă de obicei ***staționaritate în sens larg***, aceasta implicând staționaritatea momentelor de ordinul unu și doi și, deci, satisfacerea ultimelor două ecuații. Doar acest tip de procese vor fi considerate în continuare. Pentru acestea, media și varianța nu depind de timp, acestea notându-se de aceea m_x și σ_x^2. Totodată, în aceste condiții, funcțiile de moment de ordinul doi depind doar de diferența dintre momentele de timp, t_1-t_2, denumită ***timp de lag*** și notată cu τ. Dintre acestea din urmă, funcțiile de autocovarianță, care indică orice dependență liniară între valorile procesului aleator separate printr-un anumit timp de lag, au proprietăți similare. Pentru ACVF, acestea sunt următoarele:

1) $\gamma_{x(0)} = \sigma_x^2$;

2) ACVF este o funcție pară: $\gamma_{(\tau)} = \gamma_{(-\tau)}$;

3) Dacă $\gamma_{x(\tau)} = \gamma_{x(\tau+P)}$ pentru toți τ, atunci $x(t)$ este o funcție periodică cu perioada P;

4) $\left|\gamma_{x(\tau)}\right| \le \gamma_{x(0)}$.

Pentru NACVF, prima și ultima proprietate devin: $\xi_{x(0)} = 1$ și, respectiv, $\left|\xi_{x(\tau)}\right| \le 1$, această funcție având, pentru un timp de lag stabilit, semnificația unui coeficient de corelație a zgomotului care afectează semnalul la momente de timp separate prin timpul de lag considerat.

https://biblioteca-digitala.ro / https://unibuc.ro

În tabelul 7.2 se arată expresiile momentelor de ordinul unu şi doi pentru un proces aleator infinit, **înregistrat** în domeniul de timp continuu, $x(t)$, şi, respectiv, discret, $x(n)$, precum şi ale unor estimatori ai acestora obţinuţi pe baza unei înregistrări finite a semnalului, $x(n)$, $n=1, N$. Indicele x din dreapta jos a momentelor a fost omis, semnalul fiind subînţeles.

Tabelul 7.2

Expresiile momentelor de ordinul unu şi doi pentru un proces aleator staţionar şi pentru unii estimatori ai acestora

Moment	Expresii pentru momentele unui semnal		Estimatori
	Continuu	Discret	
Media	$m = \lim_{P\to\infty} \frac{1}{2P} \int_{-P}^{P} x(t)dt$	$m = \lim_{n\to\infty} \frac{1}{2N+1} \cdot \sum_{n=-N}^{N} x(n)$	$\hat{m}_N = \frac{1}{N} \sum_{n=1}^{N} x(n)$
Varianţa	$\sigma^2 = \lim_{P\to\infty} \frac{1}{2P} \int_{-P}^{P} (x(t)-m)^2 dt$	$\sigma^2 = \lim_{n\to\infty} \frac{1}{2N+1} \cdot \sum_{n=-N}^{N} (x(n)-m)^2$	$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x(n)-\hat{m}_N)^2$
ACF	$\varphi(r) = \lim_{P\to\infty} \frac{1}{2P} \int_{-P}^{P} x(t)x(t+r)dr$	$R(k) = \lim_{N\to\infty} \frac{1}{2N+1} \cdot \cdot \sum_{n=-N}^{N} x(n)x(n+k)$	$\hat{R}(k) = \frac{1}{N} \sum_{n=1}^{N-k-1} x(n)x(n+k)$ $\hat{R}'(k) = \frac{1}{N-k} \sum_{n=1}^{N-k-1} x(n)x(n+k)$
ACVF	$\gamma(\tau) = \varphi(\tau) - m^2$	$C(k) = R(k) - m^2$	
NACVF	$\xi(\tau) = \gamma(\tau)/\gamma(0)$	$\rho(k) = C(k)/C(0)$	

Deoarece, după cum s-a menţionat, tehnicile de studiu ale semnalelor stochastice prezentate necesită staţionaritatea în sens larg a proceselor aleatoare investigate, această proprietate trebuie testată statistic. În cazul semnalelor reale, modificarea unor proprietăţi complexe ca cele ale ACF sau ale funcţiilor densităţii de probabilitate sunt însoţite de modificări ale mediei şi/sau varianţei. Prin urmare, o metodă adecvată de probare a staţionarităţii în sens larg o constituie testarea schimbărilor în timp ale acestor momente. Pentru aceasta, semnalul se divide în mai multe segmente şi se compară statistic între ele mediile şi varianţele acestora.

Pentru testarea egalităţii mediilor, $\hat{m}_1$ si $\hat{m}_2$, a două segmente cu lungimile N_1 şi N_2, se utilizează statistica T definită prin:

$$T = \frac{\hat{m}_1 - \hat{m}_2}{\left[\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\frac{N_1\hat{\sigma}_1^2 + N_2\hat{\sigma}_2^2}{N_1 + N_2 - 2}\right)\right]^{1/2}}$$

https://biblioteca-digitala.ro / https://unibuc.ro

, care, în cazul staționarității semnalului, prezintă o distribuție Student cu $\nu=N_1+N_2-2$ grade de libertate, iar în cazul variațelor $\hat{\sigma}_1^2$ şi $\hat{\sigma}_2^2$ – statistica F:

$$F = \frac{(N_1/\nu_1)\hat{\sigma}_1^2}{(N_2/\nu_2)\hat{\sigma}_2^2}$$

, care urmează o distribuție F cu $\nu_1 = N_1 - 1$ şi $\nu_2 = N_2 - 1$ grade de libertate.

Segmentele trebuie să fie suficient de lungi pentru a reflecta corect proprietățile semnalului. Astfel, de exemplu, dacă semnalul prezintă componente oscilante, în fiecare segment trebuie incluse mai multe cicluri de oscilație.

Tehnicile moderne de filtrare a zgomotului conținut în semnalele analitice se bazează pe compoziția spectrală diferită a zgomotului față de cea a componentei deterministe. Analiza spectrală a oricărui semnal stochastic care prezintă staționaritate în sens larg prin utilizarea transformatei Fourier discrete, $X(f)$, a unei funcții de probă, $x(n)$, $n=1,N$, nu este posibilă deoarece $X(f)$ reprezintă o variabilă aleatoare care nici măcar nu poate fi definită deoarece funcția de probă continuă, $x(t)$, $-\infty<t<\infty$, din care s-a prelevat seria de timp nu îndeplineşte condițiile lui Dirichlet, respectiv energia semnalelor stochastice este infinită. De aceea, pentru astfel de semnale, în scopul caracterizării spectrale se utilizează, în locul funcției densitate de energie, funcția densitate de putere, $S(f)$, definită prin:

$$P_{tot} = \int_{-\infty}^{\infty} |S(f)|^2 df$$

, unde P_{tot} reprezintă puterea zgomotului. Pe baza unei serii de timp, $x(n)$, cuprinzând un număr foarte mare de valori achiziționate cu un timp de prelevare T, $S(f)$ se poate defini cu relația:

$$S(f) = \lim_{N\to\infty} E\left(\frac{X_P(f)X_P^*(f)}{(2N+1)T}\right) df$$

, în care $X_P(f)$ reprezintă transformata Fourier a unui semnal, $x_p(n)$, obținut prin trunchierea semnalului inițial cu relația:

$$x_p(n) = \begin{cases} x(n), \text{pentru } |n| \le N \\ 0, \text{pentru } |n| > N \end{cases}$$

https://biblioteca-digitala.ro / https://unibuc.ro

Funcția densitate de putere este reală $(S(f) = S^*(f))$, pară $(S(f) = S(-f))$ și pozitivă $(S(f) \geq 0)$. Toată informația referitoare la faza componentelor spectrale ale procesului aleator se pierde în funcția densitate de putere datorită primei dintre aceste proprietăți.

Estimarea clasică a spectrului de putere al zgomotului se realizează pe baza teoremelor lui Wiener – Hincin și a lui Wold [6,7] care, pentru procese în domeniu de timp continuu și, respectiv, discret afirmă că funcțiile densitate de putere și de autocorelație a semnalului aleator reprezintă perechi de transformate Fourier.

Informații detaliate despre seriile de timp ale proceselor aleatoare, inclusiv spectrul de putere, pot fi obținute prin modelarea acestora folosind conceptele de sisteme de timp discret, stabile și cauzale. În acest cadru, tratat în cursuri avansate de prelucrare a semnalelor [1], se arată că toate procesele aleatoare corespunzând unor sisteme tehnice reale pot fi reprezentate prin modele mixte, având componente autoregresive și/sau de medie alunecătoare, modele care, odată deduse prin tehnici statistice laborioase, servesc la estimarea modernă a spectrului de putere al zgomotului.

7.3.2. Tipuri de zgomot conținute în semnalele analitice

Zgomotul sistematic care afectează măsurătorile analitice poate fi, de regulă, diminuat prin adoptarea unui plan experimental potrivit. În schimb zgomotul aleator este inerent semnalului și/sau măsurării. Pentru evaluarea gradului de contaminare a unui semnal cu zgomot aleator se utilizează raportul semnal/zgomot dat de relația:

$$S/N = \frac{media\ semnalului}{deviatia\ standard\ a\ semnalului}$$

Sursele de zgomot sunt diverse: ***din mediul înconjurător*** (vibrațiile, instabilitatea rețelei electrice de alimentare, fluctuațiile de temperatură, de lumină etc), care pot fi reduse prin măsuri tehnice adecvate, ***surse fundamentale***, datorate caracterului discret al materiei și energiei, precum și fluctuațiilor tehnice, și ***zgomot de cuantificare***, care apare în procesul în procesul de măsurare în urma cuantificării valorilor semnalului de evaluat.

Sursele fundamentale de zgomot pot fi clasificate [4] după dependența zgomotului de puterea semnalului astfel:

https://biblioteca-digitala.ro / https://unibuc.ro

a) Zgomot independent de puterea semnalului (N_A=const), sau zgomot limitat de detector. Predomină la detecția radiației cu energia fotonilor ($h\nu$) mai mică decât energia fotonilor, k_BT, unde T reprezintă temperatura absolută a detectorului. Sursa acestui tip de zgomot o constituie detectorul. Zgomotul de acest tip nu depinde de puterea sursei și apare la detecția radiației IR, din domeniul microundelor și al radiofrecvențelor.

b) Zgomot proporțional cu rădăcina pătrată a puterii semnalului ($N_B \propto \propto (semnal)^{1/2}$) sau ***zgomot limitat de sursă***. El predomină când energia fotonilor este mai mare decât k_BT și când detecția (sau numărarea) fotonilor sau electronilor individuali, supusă statisticii Poisson, se face cu eficiență aproximativ unitară. Depinde de radicalul puterii semnalului și apare în spectrometria UV-Vis și în spectrometria cu particule încărcate (cu fotoelectroni, cu electroni de impact, ESCA etc).

c) Zgomot proporțional cu puterea semnalului ($N_C \propto semnal$), sau de „*fluctuație*", de „*modulare*", de „*scintilație*". Apare la detecția semnalelor foarte mici. Tipul corespunzător de zgomot din circuitele electronice se numește **zgomot 1/*ν***, deoarece mărimea sa variază invers proporțional cu frecvența.

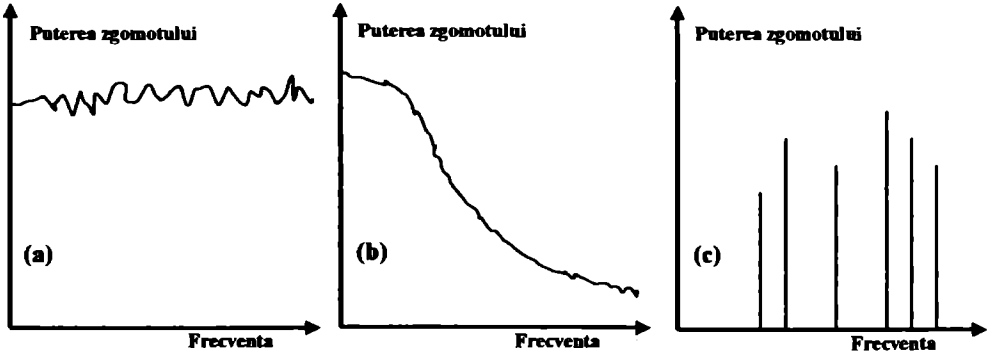


Figura 7.8. Spectrele de putere ale principalelor tipuri de zgomot. a) zgomot alb; b) zgomot de tip 1/*ν*; c) zgomot de interferență.

Zgomotul experimental ***total***, N_{total}, însumează cele trei tipuri de zgomot care, de regulă, sunt independente din punct de vedere staistic. Deci:

https://biblioteca-digitala.ro / https://unibuc.ro

$$N_{total} = (N_A^2 + N_B^2 + N_C^2)^{1/2}$$

, unde prin N^2 s-au notat varianțele zgomotului. Totuși, deși prezente simultan, în anumite tipuri de măsurători predomină un anumit tip de zgomot.

Zgomotul poate fi clasificat și în funcție de forma spectrului de putere. Trei tipuri diferite de zgomot (zgomot alb, zgomot de tip $1/\nu$ și zgomot de interferență) sunt prezentate în figura 7.8.

7.3.3. Tehnici de prelucrare a semnalelor analitice

Scopurile cele mai frecvente ale prelucrării semnalelor analitice sunt determinarea unor caracteristici fundamentale ale semnalelor (cromatograme, spectre și voltamograme), prin descrierea lor cu un număr mic de parametri; îmbunătățirea caracteristicilor metodelor de analiză: a sensibilității, aplicând tehnici de diminuare a zgomotului, și a selectivității, utilizând tehnici de restaurare a semnalelor cum ar fi cele de deconvoluție, precum și estimarea inerției de răspuns a instrumentelor analitice prin determinarea constantelor de timp ale acestora.

7.3.3.1. Determinarea constantei de timp a instrumentelor analitice

Constanta de timp reprezintă o caracteristică a analizoarelor continue legată de posibilitatea acestora de a urmări cu fidelitate fluctuațiile de concentrație. Determinarea sa se realizează analizând forma semnalelor obținute după aplicarea unui semnal cu formă cunoscută, cel mai adesea sub formă de impuls cu amplitudine constantă un timp cât mai scurt, de treaptă unitară sau cu aspect periodic sinusoidal.

Răspunsul analizorului la un impuls infinit îngust („*impulse response function*") se notează de obicei cu $h(t)$. Cunoașterea sa permite caracterizarea deplină a răspunsului analizorului, $r(t)$, deoarece acesta se obține [1] prin convoluția funcției de concentrație de la intrarea în analizor, $s(t)$, cu funcția $h(t)$:

$$r(t) = s(t) * h(t)$$

Dacă se poate înregistra răspunsul analizorului după aplicarea unui semnal treaptă de concentrație, așa cum este cazul pentru electrozii ion selectivi unde timpul de răspuns definit de IUPAC [8] este indisolubil legat de T_x, atunci constanta de timp

https://biblioteca-digitala.ro / https://unibuc.ro

se determină prin aceleaşi metode prin care se determină constanta de viteză a reacţiilor de ordinul întâi [9].

Consecinţa practică imediată a unei constante de timp diferite de zero în cazul sistemelor analitice de măsură o reprezintă deformarea semnalelor iniţiale. În general, cu cât un sistem de măsură prezintă o constantă de timp mai mare, cu atât semnalele înregistrate cu el reflectă mai puţin fidel semnalele de intrare fiind mai largi şi mai puţin înalte. Totuşi, ariile de sub curbele de variaţie ale semnalelor înregistrate în sisteme cu constante de timp diferite pentru acelaşi semnal de intrare sunt egale între ele şi, de asemenea, egale cu suprafaţa integrată a semnalului de intrare.

În general, sistemele reale monitorizate (de ex., se poate urmări variaţia de concentraţie a unui component într-un proces tehnologic sau într-un râu) prezintă şi ele o inerţie de răspuns la fluctuaţiile de concentraţie şi, de aceea, pot fi şi ele caracterizate [10] printr-o constantă de timp, $(T_x)_p$. În astfel de cazuri, între varianţele fluctuaţiilor reale de concentraţie, s_p^2, şi, respectiv, a semnalelor înregistrate cu un sistem analitic având constanta de timp T_x, σ_p^2, există relaţia:

$$\frac{s_p^2}{\sigma_p^2} = \frac{(T_X)_P}{(T_X)_P + T_x}$$

Se consideră că sistemul analitic de măsură introduce o deformare neglijabilă a semnalelor reale doar dacă constanta de timp a acestuia este de cel puţin 10 ori mai mică decât cea a procesului monitorizat. Aceasta conduce la concluzia că selectarea sistemelor de măsură pe baza valorii constantei de timp trebuie făcută în corelaţie cu procesul fizico-chimic de investigat, fapt binecunoscut spectrometriştilor care lucrează în IR în care viteza de baleiere la înregistrarea spectrului se calculează [10] în funcţie de rezoluţia dorită la înregistrarea benzilor spectrale. În plus, funcţia h(t) se determină, de obicei, înregistrând răspunsul sistemului la un semnal dreptunghiular foarte îngust şi a cărui varianţă este de cel puţin 10 ori mai mică decât a semnalului înregistrat fiind echivalentă cu acest ultim semnal.

https://biblioteca-digitala.ro / https://unibuc.ro

7.3.3.3.1.2. Sisteme analitice cu transformată Hadamard

Spectrul de absorbție al unei substanțe poate fi obținut în două moduri:

a) Se măsoară secvențial (rând pe rând) intensitatea radiației la fiecare lungime de undă (calea clasică);

b) Se utilizează o tehnică de modulare a componentelor spectrale și acestea se analizează simultan de un număr mare de ori.

Scopul principal al modulării este de a mări cantitatea de radiație care ajunge la un moment dat la detector deoarece, dacă zgomotul detectorului nu depinde de intensitatea radiației incidente, se îmbunătățește raportul *semnal/zgomot* (*S/N*).

Multiplexatea reprezintă modularea independentă și simultană a unui mare număr de componente spectrale.

Se cunosc două tehnici distincte de multiplexare:

1. Ține cont de caracterul de undă al radiației, utilizează interferometria pentru achiziția semnalelor în domeniul timpului și face apel la transformata Fourier pentru calcularea spectrului;

2. Utilizează spectrometrele cu sistem de dispersie obișnuite din care lipsește fanta de intrare și în care fanta de ieșire se înlocuiește cu o mască de codificare, care are unele zone transparente (acestea permițând trecerea radiației) și alte zone opace (reținând radiația).

Teoria matematică pe care se bazează această ultimă tehnică analitică este transformata Walsh – Hadamard, aplicată în domeniul spectrometriei pentru prima dată de către Decker și Harwit (1973) [17].

Figura 7.11 prezintă, comparativ, schemele de principiu a unui spectrometru clasic (a) și a unuia cu transformată Hadamard (b).

https://biblioteca-digitala.ro / https://unibuc.ro

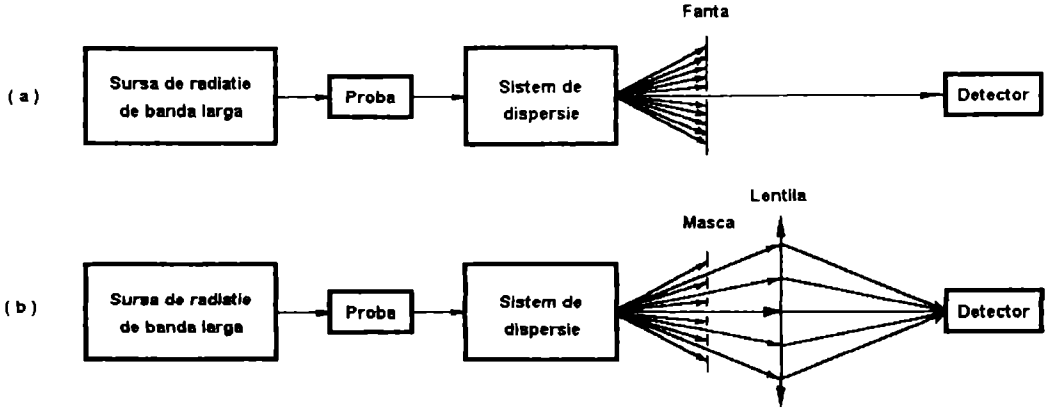


Figura 7.11. Schemele de principiu ale unui spectrometru clasic (a) şi a unuia cu transformată Hadamard (b).

În spectrometrele cu transformată Hadamard, spre deosebire de cele clasice, se interpune o mască între sistemul de dispersie şi detector. Masca se construieşte astfel încât cele mai mici deschideri ale sale să aibă aceeaşi lăţime ca şi fanta din spectrometrele clasice cu scanarea lungimii de undă, dar cu aproximativ jumătate din poziţiile posibile deschise. Alternarea zonelor închise şi deschise din calea radiaţiei este pseudoaleatoare.

Notăm cu $x_1, x_2, ..., x_N$ intensităţile transmise de probă. Când masca se află în poziţia din figura 7.12, intensitatea, y, măsurată de către detector va fi:

$$y = w_1 x_1 + w_2 x_2 + ... + w_N x_N, \tag{7.2}$$

, unde $w_n = 0$ sau 1; n=1,N, (0 – dacă fanta este deschisă, 1 – dacă este închisă).

Cu alte cuvinte, detectorul furnizează o singură observabilă (y) care depinde de N necunoscute (intensităţile spectrale independente, $x_1, x_2, ..., x_N$), conform codului din ultima ecuaţie. Situaţia este asemănătoare cu cazul A2.3 discutat în anexa 2. Pentru a putea calcula cele N valori dorite ($x_1, x_2, ..., x_N$) şi a regăsi astfel tot spectrul, ar trebui să utilizăm N măşti diferite, astfel încât fiecare să aibă un aranjament al fantelor deschise şi liniar independente de a celorlalte, problemă care ar fi foarte complicată.

Teoria matematică elaborată de Hadamard simplifică foarte mult această problemă. Astfel, se arată că, dacă N este un număr impar, atunci se poate construi o sin-

https://biblioteca-digitala.ro / https://unibuc.ro

10. D.L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte şi L. Kaufman, ***Chemometrics: A Textbook***, Elsevier, Amsterdam, 1988.

11. H. H. Rosenbrock, *Computer J.*, **3**, 174, 1960.

12. R. E. Kalman, *J. Basic Eng.*, **82**, 35, 1960.

13. S. D. Brown, *Anal. Chim. Acta*, **181**, 1, 1986.

14. R. L. Tranter, *Anal. Proc.*, **27**, 134, 1990.

15. D. L. Andrews (Editor), ***Perspectives in Modern Chemical Spectroscopy***, Springer-Verlag, 1990.

16. R. E. McKean şi D. J. Curran, *Talanta*, **39**, 319, 1992.

17. J. Decker şi M. Harwit, *Proc. Applications of Walsh Functions*, 1973. (AD-763-000).

18. A. Economu şi P. R. Fielden, *Anal Chim. Acta*, **305**, 165, 1995.

19. A. Savitzki şi M. J. E. Golay, *Anal. Chem.*, **36**, 1627, 1964.

20. S. Wolfram, ***Mathematica^{TM}. A System for Doing Mathematics by Computer***, Addison-Wesley, New York, 1988.

21. A. Aoki; T. Matsue şi I. Uchida, *Anal. Chem.*, **64**, 44, 1992.

22. I. Gh. Şabac; P. Cocârlan; O. Stănăşilă şi A. Topală, ***Matematici Speciale***, vol. II, E.D.P., Bucureşti, 1983.

23. F. Harris, *Proc. IEEE*, **66**, 51, 1978.

24. J. Ruzicka şi E. H. Hansen, ***Flow Injection Analysis***, Wiley, New York, 2nd Edition, 1988.

25. V. David, J. Martinez Calatayud şi J. V. Garcia Mateo, *Anal. Lett.*, **34**, 1553, 2001.

https://biblioteca-digitala.ro / https://unibuc.ro

ANEXA 1

Analiza spectrală a semnalelor deterministe în domeniu de timp continuu

A1.1. Funcții ortogonale

Seturile de funcții ortogonale, $\{\Phi_m(t)\}$, pot fi folosite pentru descrierea semnalelor continue într-un domeniu de timp finit cu durata P. Aceste semnale pot fi scrise, cu ajutorul unor astfel de funcții, ca serii de forma:

$$f(t) = A_0\Phi_0(t) + A_1\Phi_1(t) + \ldots + A_m\Phi_m(t) + \ldots = \sum_{m=0}^{\infty} A_m\Phi_m(t) \qquad \text{(A1)}$$

Setul de funcții este ortogonal dacă:

$$\int_0^P \Phi_m(t)\Phi_n(t)dt = \begin{cases} \lambda_n, \text{ pentru : } m = n \\ 0, \text{ pentru : } m \neq n \end{cases}$$

iar domeniile uzuale pentru m sunt $0 < m < \infty$ (ca şi mai sus) şi, respectiv, $-\infty < m < \infty$.

Coeficienții seriei A1 se determină utilizând principiul celor mai mici pătrate ale erorilor. Definind o aproximație a funcției $f(t)$ folosind un număr finit de termeni, $M + 1$, prin:

$$\hat{f}(t) = A_0\Phi_0(t) + A_1\Phi_1(t) + \ldots + A_M\Phi_M(t)$$

, eroarea pătratică a aproximării este:

$$E_M = \int_0^P \left(f(t) - \hat{f}(t)\right)^2 dt = \int_0^P \left(f(t) - \sum_{m=0}^{M} A_M\Phi_M(t)\right)^2 dt =$$

$$= \int_0^P \left[f(t)^2 - 2f(t)\sum_{m=0}^{M} A_M\Phi_M(t) + \left(\sum_{m=0}^{M} A_M\Phi_m(t)\right)^2\right] dt$$

Derivata parțială în raport cu un anumit coeficient, A_m, trebuie să fie zero:

$$\frac{\partial E_M}{\partial A_m} = -2\int_0^P f(t)\Phi_m(t)dt + 2\int_0^P \left(\sum_{n=0}^{M} A_n\Phi_n(t)\right)\Phi_m(t)dt = 0$$

Rearanjând ecuația precedentă şi utilizând condițiile de ortogonalitate, se ajunge la:

https://biblioteca-digitala.ro / https://unibuc.ro

$$\Delta f = f_0 = \lim_{P \to \infty} \frac{1}{P} = 0$$

Pentru funcția de timp, $x(t)$, rezultă următoarele relații:

$$x(t) = \int_{-\infty}^{\infty} X(f) \exp(j2\pi ft) df = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(w) \exp(jwt) dw \qquad \text{(A7)}$$

De asemenea:

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft) dt \qquad \text{(A8)}$$

, sau:

$$X(w) = \int_{-\infty}^{\infty} x(t) \exp(-jwt) dt \qquad \text{(A9)}$$

Ecuațiile (A8) și (A9) reprezintă expresiile CFT directe iar cu (A7) se calculează CFT inversă. $X(t)$ și $x(f)$ sau $x(t)$ și $X(w)$ reprezintă perechi de transformate Fourier, fapt care, de obicei, se notează pe scurt astfel: $x(t) \Leftrightarrow X(f)$ sau $x(t) \Leftrightarrow X(w)$.

Proprietățile transformatei Fourier sunt similare cu cele ale coeficienților seriilor Fourier. Se verifică ușor că părțile reală și imaginară ale transformatei Fourier sunt funcții pare și, respectiv, impare, adică: $\mathrm{Re}[X(f)] = \mathrm{Re}[X(-f)]$ și $\mathrm{Im}[X(f)] = -\mathrm{Im}[X(-f)]$, ceea ce se mai poate exprima prin relația:

$$X(f) = X^*(-f) \qquad \text{(A10)}$$

În formă polară, se poate scrie:

$$X(f) = |X(f)| \exp[\theta(f)]$$

, unde:

$$|X(f) = \sqrt{(\mathrm{Re}^2[X(f)] + \mathrm{Im}^2[X(f)])}|$$

, si:

$$\theta(f) = -\arctan \frac{\mathrm{Im}[X(f)]}{\mathrm{Re}[X(f)]}$$

$|X(f)|$ este o funție pară iar $\theta(f)$ – una impară.

Energia semnalelor aperiodice se definește prin:

https://biblioteca-digitala.ro / https://unibuc.ro

$$E_{tot} = \int_{-\infty}^{\infty} x^2(t)dt = \int_{-\infty}^{\infty} x(t)\left(\int_{-\infty}^{\infty} X(f)\exp(j2\pi ft)df \right)dt$$

Cu alte cuvinte, energia reprezintă aria de sub pătratul spectrului magnitudinilor. Funcția $|X(f)|^2$ este energia pe unitatea de frecvență și se mai numește **spectrul densității de energie**.

ANEXA 2

Planuri experimentale de cântărire

Yates (1935) a arătat că prin cântărirea mai multor obiecte în grup este posibil să se determine mai precis masa obiectelor individuale. Pentru a demonstra acest lucru, presupunem că avem de cântărit patru obiecte cu masele necunoscute x_1, x_2, x_3, x_4, și că utilizăm în acest scop o balanță analitică cu două talere exactă a cărei precizie, cuantificată prin varianța erorilor experimentale, σ^2, nu depinde de masa care se află pe talere. Considerăm trei planuri experimentale diferite de cântărire:

A2.1. Cazul clasic (al cântăririlor secvențiale). Pe unul din talerele balanței se pun pe rând cele patru obiecte. Balanța se echilibrează punând greutăți pe celălalt taler. Fie aceste greutăți y_1, y_2, y_3 și y_4. Datorită erorilor experimentale care afectează cântăririle, între masele reale, x_i, și cele determinate experimental, y_i, există relațiile:

$$y_i = x_i + e_i, \quad i = 1, 2, 3 \text{ si } 4$$

Pe baza cântăririlor efectuate, valorile estimate, $\hat{x}_i$, pentru masele obiectelor individuale sunt:

$$\hat{x} = y_i, \quad i = 1, 2, 3 \text{ si } 4$$

Valorile de așteptare ale estimațiilor sunt egale cu masa exactă a obiectelor cântărite:

$$E[\hat{x}] = x_i, \quad i = 1, 2, 3 \text{ si } 4$$

https://biblioteca-digitala.ro / https://unibuc.ro

, ceea ce este în acord cu faptul că balanța este exactă. Aşadar, deşi cântăririle individuale pot fi afectate de erori întâmplătoare, media unui mare număr de cântăriri repetate ale unui obiect nu diferă semnificativ de valoarea reală a masei obiectului respectiv.

Conform ipotezelor referitoare la precizia balanței utilizate, valorile de aşteptare ale varianțelor erorilor sunt aceleaşi, σ^2:

$$E\left[(\hat{x}_i - x_i)^2\right] = E\left[e_i^2\right], \quad i = 1, 2, 3 \text{ si } 4$$

Aceasta semnifică faptul că, deşi nu se poate afirma cu certitudine care este mărimea erorii care afectează valoarea masei unui obiect determinată printr-un singur experiment de cântărire, e_i, cel mai probabil aceasta este apropiată, în valoare absolută, de deviația standard, σ.

O cale de micşorare a erorii care afectează determinarea masei unui obiect individual o constituie cântărirea repetată a obiectului şi medierea rezultatelor obținute. Eficiența acestei metode este însă limitată de faptul că sunt necesare un număr foarte mare de determinări individuale pentru o îmbunătățire semnificativă a preciziei. Astfel, din modul de calcul al varianței unei sume de variabile aleatoare, rezultă că, dacă erorile care afectează cântăririle individuale sunt independente între ele, pentru a micşora de n ori deviația standard a rezultatului, cântărirea trebuie repetată de n^2 ori. De aceea, s-au căutat şi alte metode mai eficiente de îmbunătățire a preciziei cântăririlor. Ele se aplică însă doar atunci când trebuie determinate masele mai multor obiecte, fiind cu atât mai eficiente cu cât numărul obiectelor de cântărit este mai mare.

A2.2. Cazul cântăririlor în grup punând obiectele de măsurat pe ambele talere. În prima cântărire, toate obiectele se aşează pe un singur taler iar în următoarele trei – se combină pe rând obiectele între ele pentru a forma două grupuri de câte două obiecte. Considerând obiectele numerotate, un mod de grupare posibil este: întâi 1 cu 3 şi 2 cu 4, apoi 1 cu 2 şi 3 cu 4 şi, în final, 1 cu 4 şi 2 cu 3. Simultan, grupurile de obiecte se aşează pe talerele balanței (câte un grup pe fiecare taler) iar diferența de masă se compensează adăugând greutăți pe talerul cu obiectele mai uşoare. În acest caz, între masele greutăților utilizate la echilibrarea balanței, y_i, şi cele ale obiectelor individuale, x_i, există relațiile următoare:

https://biblioteca-digitala.ro / https://unibuc.ro

$$y_1 = x_1 + x_2 + x_3 + x_4 + e_1$$
$$y_2 = x_1 - x_2 + x_3 - x_4 + e_2$$
$$y_3 = x_1 + x_2 - x_3 - x_4 + e_3$$
$$y_4 = x_1 - x_2 - x_3 + x_4 + e_4$$

Estimațiile maselor obiectelor, obținute prin rezolvarea acestui sistem de ecuații, sunt următoarele:

$$\hat{x}_1 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4) = x_1 + \frac{1}{4}(e_1 + e_2 + e_3 + e_4)$$
$$\hat{x}_2 = \frac{1}{4}(y_1 - y_2 + y_3 - y_4) = x_2 + \frac{1}{4}(e_1 - e_2 + e_3 - e_4)$$
$$\hat{x}_3 = \frac{1}{4}(y_1 + y_2 - y_3 - y_4) = x_3 + \frac{1}{4}(e_1 + e_2 - e_3 - e_4)$$
$$\hat{x}_4 = \frac{1}{4}(y_1 - y_2 - y_3 + y_4) = x_4 + \frac{1}{4}(e_1 - e_2 - e_3 + e_4)$$

, iar pentru varianțele erorilor care afectează aceste rezultate, se obține:

$$E\left[(\hat{x}_i - x_i)^2\right] = \frac{4\sigma^2}{16} = \frac{\sigma^2}{4}, \quad i = 1, 2, 3 \text{ şi } 4$$

, deoarece varianța sumei sau diferenței unor variabile aleatoare independente este egală cu suma varianțelor variabilelor aleatoare individuale. Exemplificând deducerea ultimei relații pentru $i = 1$, rezultă într-adevăr că:

$$E\left[(\hat{x}_1 - x_1)^2\right] = E\left[\left\{\frac{1}{4}(e_1 + e_2 + e_3 + e_4)\right\}^2\right] = \frac{1}{16}E\left[(e_1 + e_2 + e_3 + e_4)^2\right] = \frac{4\sigma^2}{16} = \frac{\sigma^2}{4}$$

Se observă că, utilizând acest plan de cântărire, se reduce de patru ori varianța erorilor de măsură iar raportul semnal/zgomot creşte de două ori față de cazul în care obiectele sunt cântărite individual.

A2.3. Cazul cântăririlor în grup punând obiectele pe un singur taler. În acest caz, vom considera că obiectele se grupează astfel: mai întâi obiectele 2, 3, şi 4 şi, după aceea, pe rând, obiectul 1 cu fiecare din celelalte trei, şi că, la fel ca şi în cazul A2.2, grupurile de obiecte se aşează pe un singur taler, greutatea lor fiind compensată cu greutățile $y_i, (i = 1,4)$. Se poate acum scrie:

$$y_1 = x_2 + x_3 + x_4 + e_1$$
$$y_2 = x_1 + x_2 + e_2$$

https://biblioteca-digitala.ro / https://unibuc.ro

$$y_3 = x_1 + x_3 + e_3$$
$$y_4 = x_1 + x_4 + e_4$$

, de unde se obțin următoarele valori de estimație ale obiectelor de cântărit:

$$\hat{x}_1 = \frac{1}{3}(y_1 + y_2 + y_3 + y_4) = x_1 + \frac{1}{3}(e_1 + e_2 + e_3 + e_4)$$
$$\hat{x}_2 = \frac{1}{3}(2y_2 + y_1 - y_3 - y_4) = x_2 + \frac{1}{3}(2e_2 + e_1 - e_3 - e_4)$$
$$\hat{x}_3 = \frac{1}{3}(2y_3 + y_1 - y_2 - y_4) = x_2 + \frac{1}{3}(2e_3 + e_1 - e_2 - e_4)$$
$$\hat{x}_4 = \frac{1}{3}(2y_4 - y_1 - y_2 + y_3) = x_2 + \frac{1}{3}(2e_4 - e_1 - e_2 + e_3)$$

, şi următoarele valori de aşteptare pentru varianțele acestora:

$$E\left[(\hat{x}_1 - x_1)^2\right] = \frac{4\sigma^2}{9}$$
$$E\left[(\hat{x}_i - x_i)^2\right] = \frac{7\sigma^2}{9},\ i = 2-4$$

În concluzie, şi în acest caz se îmbunătățeşte raportul semnal/zgomot față de cazul A2.1, dar într-o măsură mai mică decât în cazul A2.2.

A2.4. Cazul general de cantarire în grup. În general, dacă avem p necunoscute, $(x_1, x_2, ..., x_p)$, şi N $(N > p)$ măsurători, $(y_1, y_2, ..., y_N)$, afectate de erorile $(e_1, e_2, ..., e_N)$, putem scrie setul de ecuații:

$$y_i = w_{i1}x_1 + w_{i2}x_2 + ... + w_{iN}x_N, \qquad i = 1, N$$

, care, în notații matriceale, poate fi scris astfel:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{e}$$

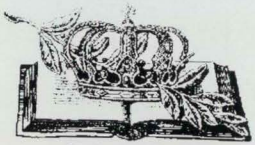
, unde $\boldsymbol{W}$ reprezintă o matrice cu N rânduri şi p coloane. În cazul experimentelor de cântărire, coeficienții $w_{ij}, (i = 1, N \text{ si } j = 1, p)$ pot lua doar valorile 0, 1 şi –1 iar specificarea lor completă se numeşte *„plan experimental de cântărire"*

VERIFICAT 2017

https://biblioteca-digitala.ro / https://unibuc.ro

DATA RESTITUIRII		
17 FEB. 2005		
16 DEC. 2010		
17 DEC. 2010		

BIBLIOTECA CENTRALA
UNIVERSITARA „CAROL I"



DE SPIRITU ET ANIMA

https://biblioteca-digitala.ro / https://unibuc.ro

140.000 lei.

ISBN 973-575-917-9

Lei 140000

https://biblioteca-digitala.ro / https://unibuc.ro